# SANDIA REPORT

SAND 2005-7989
Unlimited Release
Printed December 2005

# Reverse Engineering Biological Networks: Applications in Immune Responses to Bio-Toxins

Jean-Loup Faulon, Zhaoduo Zhang, Anthony Martino, Jerilyn A. Timlin, David M. Haaland, Edward V. Thomas, Michael B. Sinclair, Shawn Martin, George Davidson, Elebeoba May, and Alex Slepoy

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.



Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865)576-8401
Facsimile: (865)576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.doe.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800)553-6847
Facsimile: (703)605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND 2005-7989
Unlimited Release
Printed December 2005

# Reverse Engineering Biological Networks: Applications in Immune Responses to Bio-Toxins

Jean-Loup Faulon
Computational Biosciences Department
Zhaoduo Zhang
Biosystem Research department
Anthony Martino, Jerilyn A. Timlin, David M. Haaland,
Biomolecular Analysis & Imaging Department
Edward V. Thomas
Independent Surveillance Assessment & Statistics Department
Michael B. Sinclair
Microsystem Materials Department
Shawn Martin, George Davidson, and Elebeoba May
Computational Biology Department
Alex Slepoy
Multiscale Computational Material Methods
Sandia National Laboratories
PO Box 5800
Albuquerque NM, 87185

ABSTRACT

Our aim is to determine the network of events, or the regulatory network, that defines an immune response to a bio-toxin. As a model system, we are studying T cell regulatory network triggered through tyrosine kinase receptor activation using a combination of pathway stimulation and time-series microarray experiments. Our approach is composed of five steps 1) microarray experiments and data error analysis, 2) data clustering, 3) data smoothing and discretization, 4) network reverse engineering, and 5) network dynamics analysis and fingerprint identification. The technological outcome of this study is a suite of experimental protocols and computational tools that reverse engineer regulatory networks provided gene expression data. The practical biological outcome of this work is an immune response fingerprint in terms of gene expression levels.

Inferring regulatory networks from microarray data is a new field of investigation that is no more than five years old. To the best of our knowledge, this work is the first attempt that integrates experiments, error analyses, data clustering, inference, and network analysis to solve a practical problem. Our systematic approach of counting, enumeration, and sampling networks matching experimental data is new to the field of network reverse engineering. The resulting mathematical analyses and computational tools lead to new results on their own and should be useful to others who analyze and infer networks.

ACKNOWLEDGMENT

This work was funded by Sandia Laboratory Directed Research and Development. Some of the work was also funded by the US Department of Energy's Genomics: GTL program (www.doegenomestolife.org) under project, "Carbon Sequestration in Synechococcus Sp.: From Molecular Machines to Hierarchical Modeling," (www.genomes-to-life.org). Sandia is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

# Contents

# List of Figures



# List of Tables

# Chapter 1

# Overview

We present an experimental/computation protocol to infer networks of regulation between genes that defines an immune response to a bio-toxin. We anticipate that by inferring a regulatory network we will be able to isolate a handful of specific regulatory elements that fingerprint a response.

As a model system, we are studying T cell regulatory network triggered through tyrosine kinase receptor activation using a combination of pathway stimulation and time-series microarray experiments. T cell recognition of a bio-toxin (i.e., a foreign antigen) initiates a network of signals, protein pathways, and gene inductions that regulate cellular growth and proliferation [1]. The initial signal induced by a foreign antigen leads to, among other things, the induction of interleukin-2 (IL-2). IL-2 is expressed and released as a soluble factor that in an autocrine fashion stimulates the very cell from which it was released. IL-2 recognition forms a second signal occurring in series. The combined result of the two signals is the activation of protein networks leading to the up-regulation of genes responsible for growth and proliferation. Protein networks and genes are regulated by either one of the signals or by some combination of both signals. Despite years of research, only a handful of partial protein networks and induced genes are known to exist uniquely downstream of one signal or the other. Even fewer shared protein pathways and genes are known, and there is no understanding of how the signals co-regulate responses. Results are limited by trial and error experimental designs that focus on single proteins or genes.

We propose to utilize recently developed high-throughput experiments (DNA microrray) couple with computational reverse engineering techniques (gene network inference) to discover the regulatory processes in T cell signaling that lead to growth and proliferation. Developing large data sets and a fuller understanding of the contextual signals will facilitate inference of regulatory networks that, in turn, will provide a better understanding of T cell response to bio-toxin.

As illustrated in Figure 1.1, our approach is composed of five steps 1) microarray experiments and data error analysis, 2) data clustering, 3) data smoothing and discretization, 4) reverse engineering, and 5) network dynamics analysis and fingerprint identification.

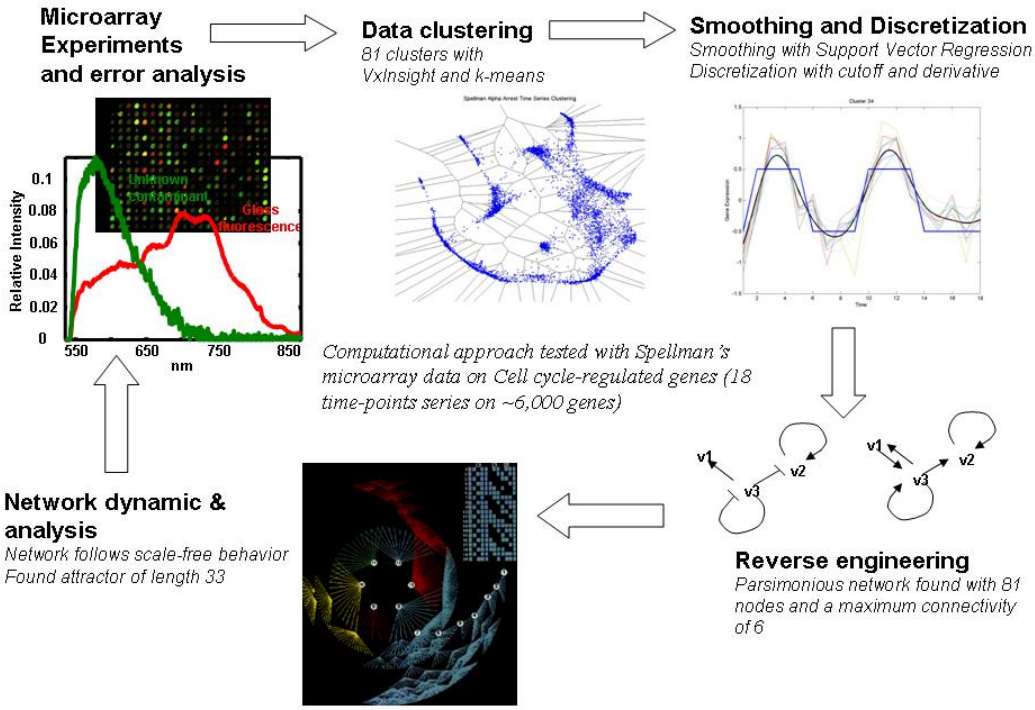


Figure 1.1. Overall experimental/computational protocol to reverse engineer regulation between genes. The figure is illustrated with Yeast cell cycle microarray data [2].

The five steps illustrated above are briefly outlined next and described in further detail in the remainder of the report. In particular, we discuss in Chapter 2 microarray and error analysis. In Chapter 3 we describe clustering, smoothing and discretization, and reverse engineering using as an example Yeast cell cycle microarray data [2]. In Chapter 4 we present some findings related to network dynamics analysis, and in Chapter 5 we illustrate the entire experimental/computational protocol using IL-2 starved and IL-2 stimulated T cell microarray data. Furthermore, in Chapter 5 we determine a list of genes representative of the T cell immune response to IL-2 stimulation.

## 1.1. MICROARRAY EXPERIMENTS AND ERROR ANALYSIS

Microarray technology has made it possible to detect and quantify the induction of thousands of genes in a single experiment. Gene microarrays consist of small pieces of DNA representing single genes that are placed systematically and with known addresses on a chip. Cells, in our case mouse T cells, are cultured under specific experimental conditions and bulked up to a critical concentration. The cells are lysed and all of the cellular mRNA is harvested. Each mRNA molecule, representing an induced gene within the cells, is allowed to hybridize (find its complement) to DNA on the chip. Positives (detected by radiography or fluorescence) indicate the gene at a particular address has been induced and at what concentration.

In order to understand more fully T cell signaling and to generate large data sets to infer regulatory networks, in Chapter 5, we measure gene induction events using array technology over a time course combined with specific conditions designed to stimulate various protein pathways. Time dependence is critical in detecting clusters of genes with similar kinetics thus implying common regulatory pathways.

It is well known that microarray data involve a large number of error-prone steps resulting in a high level of noise. As further described in Chapter 2, to improve curve resolution and the accuracy and dynamic range obtained from microarray fluorescence experiments, we use a hyperspectral microarray scanner and a newly developed multivariate curve resolution algorithm. These new data analysis methods allow us to model fluorescence dyes and all sources of background emission at each pixel. Therefore, we do not require the inaccurate step of subtracting a background next to the spot from the spot fluorescence intensity. Our normalization of the fluorescence data is also not subject to the many error sources present in microarray data collected with current commercial scanners.

## 1.2. MICROARRAY DATA CLUSTERING AND DISCRETIZATION

Clustering of expression profiles is an important step prior to reverse engineering networks as it reduces the number of variables. Indeed, there is no reason to distinguish two genes if one cannot distinguish their expression profiles from the data. Aside from inferring networks, data clustering can also be used to determine co-regulated genes directly from the microarray experiments. In this study use k-mean as a clustering technique and Support Vector Regression (SVR) for discretization. Clustering and discretization are further discussed in Chapters 3 and 5 and illustrated with Yeast cell cycle and IL-2 stimulated T cell microarray data.

## 1.3. NETWORK REVERSE ENGINEERING

We construct a coarse-scale model of the network of regulatory interactions between gene clusters determined in the previous step. Precisely, we study the four following problems, which are further detailed in Chapter 5: 1) INFER: Searches for a network matching the expression profiles and thus verifies the consistency of the data. 2) COUNT: Count the number of networks matching the expression profiles. This routine

will provide a measure of the degeneracy of a data set. 3) ENUMERATE: Enumerates all the networks matching the expression profiles. 4) SAMPLE: Samples networks at random matching the expression profiles. The above algorithms are used to infer networks for Yeast cell cycle (Chapter 3) and IL-2 stimulated T cell (Chapter 5).

## 1.4. NETWORK DYNAMIC ANALYSIS AND FINGERPRINT IDENTIFICATION.

Network dynamic analysis is performed to 1) verify that the expression profiles given as input can indeed be reproduced by the inferred networks 2) to expand the dynamics beyond the times series that were provided as input, and 3) to predict expression profiles under initial conditions different than those provided. Of particular interest is to compute the steady state or equilibrium dynamics of the networks, these steady states should determine the final response of the networks under the initial conditions they are subjected to. With Boolean networks and Boolean dynamics, these steady states are named attractors [3]. An attractor is a cyclic pattern of expression profiles in which the dynamics is drawn too. That cyclic pattern and the corresponding genes represent the fingerprint response of the network. Techniques to compute and analyze attractors are presented in Chapters 4 and 5. Furthermore, in Chapter 4 we analyze the robustness of inferred networks by computing how much a network can be perturbed without changing its steady state dynamics. We also compare in this Chapter the robustness of gene regulatory networks versus random networks.

Inferring regulatory networks from microarray data is a new field of investigation that is no more than five years old. To the best of our knowledge, this work is the first attempt that integrates experiments, error analyses, data clustering, inference, and network analysis to solve a practical problem. The systematic approach of counting, enumerating, and sampling is new to the field of network reverse engineering. The resulting

mathematical analyses and computational tools should lead to new results on their own and should be useful to others who analyze and infer networks.

## 1.5. REFERENCES

[1] A. Martino *et al.* Journal of Immunology, 166 (2001), 1723.

[2] P. T. Spellman *et al*. Molecular Biology of the Cell, 9 (1998), 3273.

[3] S.A. Kauffman. J. Theor. Biol. 22 (1969), 437.

# Chapter 2

# Hyperspectral imaging of biological targets: The difference a high resolution spectral dimension and multivariate analysis can make

This work was done in collaboration with M. Juanita Martinez[b], Monica Manginell[a], Susan M. Brozik[a], and Margaret Werner-Washburne[b]

[a]Sandia National Laboratories, Albuquerque, NM 87111
[b]Department of Biology, University of New Mexico, Albuquerque, NM 87131

**Portions reprinted, with permission, from *"Hyperspectral imaging of biological targets: The difference a high resolution spectral dimension and multivariate analysis can make."* Timlin, J.A.; Sinclair, M.B.; Haaland, D.M.; Martinez, M.J.; Manginell, M.; Brozik S.M.; Guzowski, J.F.; Werner-Washburne, M. Biomedical Imaging: Macro to Nano, 2004. IEEE International Symposium on 15-18 April 2004 Page(s): 1529- 1532.** 

## ABSTRACT

*Hyperspectral imaging coupled with multivariate data analysis is a powerful new tool for understanding complex biological and biomedical samples. The advantages and drawbacks of adding a spectral dimension and multivariate data analysis to optical microscopy for biological interrogation will be demonstrated with applications of DNA microarrays and live cell imaging. These data are selected to present the type of impact hyperspectral imaging can have in biomedical science. Images are acquired using our state-of-the-art hyperspectral imaging system and multivariate data analysis is used to*

*extract pure component spectra and corresponding independent concentration maps of all fluorescent species. In most cases the data analysis algorithms are successful with little or no information given* a priori *and generate images that are free of the influences of spectral crosstalk, cellular autofluorescence, and other background emissions that often plague traditional fluorescence microscopy.*

## 2.1. INTRODUCTION AND BACKGROUND

Fluorescence microscopy is an important tool for the diagnosis of disease, the mapping of proteins and genes, the classification of cells and tissues, and the identification of cellular interactions and pathways. Traditional fluorescence microscopes are filter-based instruments that pass all photons emitted from a sample within a specific wavelength range (band pass) to the detector. Multicolor (multicomponent) images can be created through the use of independent scans, careful choice of filter sets, and an even more critical choice of fluorophores to minimize spectral overlap between the channels. In contrast, a hyperspectral imaging microscope collects an entire emission spectrum at each image pixel. This approach facilitates simultaneous, single-scan imaging of spectrally overlapped emissions. Previous research using hyperspectral imaging has outlined the potential of the technique for applications areas such as flow cytometry and histology.[1] Recently a module to add multispectral imaging capabilities to a popular confocal microscope has been introduced.[2]

We have developed and characterized a hyperspectral imaging system optimized for scanning printed DNA microarrays.[3] This system operates as a push-broom style line imager and is capable of collecting fluorescence emission spectra (3 nm spectral resolution) from 400-900 nm. The system was designed with low spatial resolution (10 μm) for scanning microarrays, but can accommodate more demanding biology applications with minor modifications to the optics. For the cell-based application presented here, modifications were made to achieve 2.5 μm spatial resolution. While acceptable for this preliminary investigation, this resolution is fairly coarse for cellular imaging and future hardware modifications will allow imaging at the diffraction limit.

The hyperspectral imaging experiments described in this communication produce a data cube with one spectral and two spatial dimensions. The advantages of the spectral dimension of information depend on the experiment but can include: 1. increased multiplexing due to relief of the requirement that the labels be non-overlapping, 2. the ability to identify and model unwanted emissions such as background emissions, cellular autofluorescence, and contaminants, and thus removing their effect on the species of interest, 3. increased reliability, precision and dynamic range due to properties of the spectral data. However, it is critical to note that none of these advantages can be fully realized without the use of multivariate data analysis to extract the latent relationships within the complex data cube. As the utility of multivariate analysis is realized analysis techniques such as classical least squares, linear unmixing, and SIMPLISMA are increasing in popularity. We use principal component analysis followed by multivariate curve resolution (a constrained, iterative alternating least squares technique, optimized in our laboratory for hyperspectral image analysis on desktop PC's) to extract pure component spectra and corresponding independent concentration maps from our hyperspectral images. The primary advantage of multivariate curve resolution over other multivariate algorithms is its ability to determine spectral components with little or no information a priori. Algorithm details are given elsewhere and are not the focus of the current work.[4, 5]

The work described in this communication serves to outline the potential hyperspectral imaging and multivariate data analyses have for biomedical imaging. The three examples presented were chosen because they describe the value of adding a high resolution spectral dimension to biological microscopy for increased multiplexing, identification of unwanted emissions, and reduction of the effects of extraneous emissions.

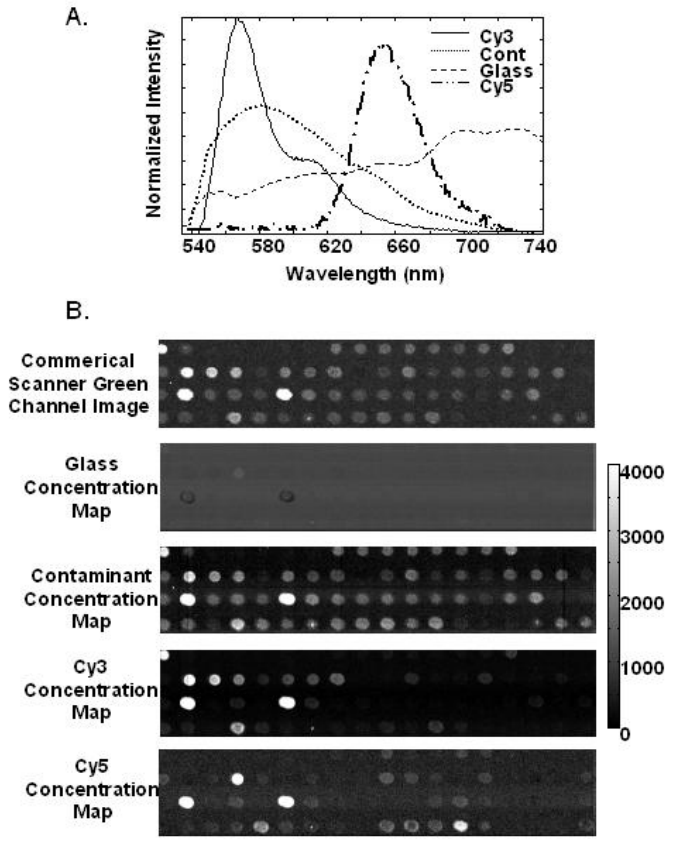


**Figure 2.1 A. Extracted pure component spectra from hyperspectral image of microarray with contaminating fluorescence. B. Corresponding concentration maps. An appropriately scaled green channel image from a commercial scanner is shown to compare the degree of contamination. Spots are ~ 200 μm in diameter.**

## 2.2. RESULTS AND DISCUSSION

### 2.2.1. Identifying emissions on printed DNA microarrays

Researchers rely on DNA microarrays to provide relative gene expression information in a high throughput format. A typical microarray experiment consists of fluorescent probe DNA (usually Cy5 and Cy3) from two sources (e.g. control and tumor cells) that is allowed to hybridize to a spotted gene library. Red and green emission images are collected and after correcting for background emissions, R/G ratios are constructed to determine differentially expressed genes. Ideally the only emission from within a printed

DNA spot should be the labeled DNA and glass and the only emission outside the spot should be the glass substrate. However there is evidence in the literature for additional, contaminating emissions.[6, 7,8]

In recent work with collaborators at the University of New Mexico we have used our optimized hyperspectral microarray scanner and multivariate data analysis procedures to identify and quantitate a contaminant confounded in the green channel images of commercial microarray scanners in the presence of the green and red labeled DNA.[8] This spot-specific, variable contaminant is particularly detrimental because it invalidates the background correction and normalization procedures used in microarray analysis. Figure 2.1 illustrates the spectral components and corresponding concentration maps resulting from multivariate analysis of a hyperspectral image of a DNA microarray exhibiting contaminating fluorescence. The hyperspectral image was collected at 10 μm spatial resolution using a 532 nm laser for excitation of all fluorophores. The extracted concentration maps clearly show that the contaminant dominates the signal in the commercial image, leading to erroneous results. In this particular case 75% of the green channel intensities from the commercial scanner are in error by a factor of 2 or more. Although this is just one example of our extensive work with microarrays, it reveals the power of hyperspectral imaging for identifying and correcting for unexpected emission sources in fluorescence microscopy and imaging.

### 2.2.2. Performing quality control on microarrays used for inferring regulatory networks

As part of our recent collaboration with the authors of this SAND report we applied our hyperspectral imaging techniques in combination with rigorous statistical design of experiments to assess the quality of microarray data for inferring regulatory networks. The arrays were examined for extraneous emissions in the green channel. The presence of these emissions would reduce the accuracy and reliability of the microarray data and could ultimately lead to in accurate conclusions of network pathways if measures are not taken to correct the data. Hyperspectral scanning experiments were done to assess the print and hybridization quality of a murine genome array printed and hybridized by B. Griffith at UNM and then

subsequently, the same murine arrays hybridized at Sandia National Laboratories. Areas on a total of 18 slides (10 hybridized for this project and 8 hybridized for other projects) were scanned using the hyperspectral scanner and analyzed using multivariate curve resolution. The background levels (dye outside the spots) were found to be very good, but the data analysis did indicate the presence of a spot-specific, non-uniform contaminant in the green channel similar to the contaminant we have seen previously in microarrays from other labs. The level of contaminant was low (the UNM lab printing the slides has taken the known steps to minimize the contaminant), but in many cases the signal from the hybridized dye was also very low and thus the majority of the signal within a spot was often largely contaminant. It was recommended on the basis of this discovery that it would be beneficial to explore ways to increase the signal from the hybridized DNA such as by increasing the amount of starting RNA. Figure 2.2 shows the hyperspectral imaging results from one of the mouse microarrays scanned as part of the reverse engineering of networks study. It is clear from these results that the hyperspectral imager provides the power to identify and correct microarray data for the presence of extraneous emissions.

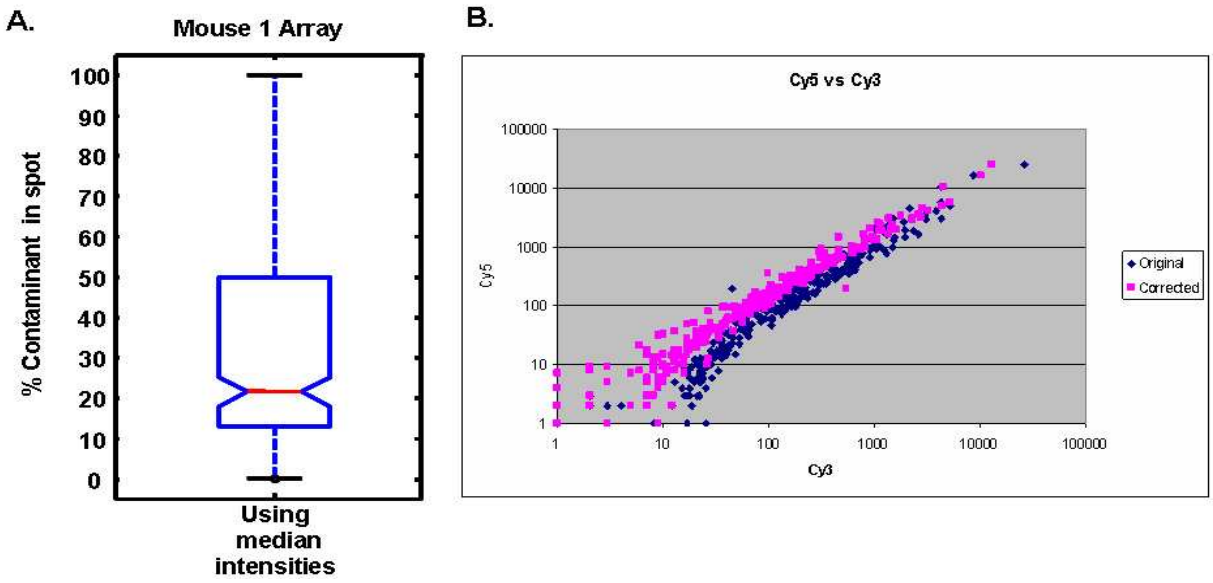


**Figure 2.2: Hyperspectral imaging results from a mouse microarray. A. Hyperspectral scanning found generally low levels of green contaminant present (20% contaminant on average). B. The harmful effect of this green contaminant on the Cy5/Cy3 ratio is shown as a skew to the green channel in the uncorrected array data. Data was corrected by using the independent Cy5 and Cy3 concentration values from the hyperspectral image analysis.**

### 2.2.3. Simultaneous imaging of multicolor fluorescent proteins in live cells

Multicolor fluorescent protein experiments could be ideal for identifying multiple proteins and gene expression in vivo, but are often complicated by spectral crosstalk between fluorophores and cellular autofluorescence.[9] These factors limit the number of useful fluorescent protein markers. Researchers have attacked this limitation by modifying the fluorescent protein structure to create variants with more desirable excitation/emission properties[10] and designing specialized optical filters, but this approach does not entirely solve the problem.

Three strains of yeast cells were genetically engineered to express cyan fluorescent protein (CFP), green fluorescent protein (GFP), or yellow fluorescent protein (YFP) when exposed to galactose. These strains were grown in media with galactose for 4 hrs to induce protein expression. Equal volumes of each of the three colors of cells were mixed and placed on a quartz slide under a quartz coverslip for imaging. The sample was excited using 488 nm light and fluorescence emission spectra were collected over a 600μm square area in a single 60s scan. With appropriate constraints the multivariate analysis was able to reduce the ~58000 spectra dataset to a smaller set of pure components— CFP, GFP, YFP, and cellular autofluorescence. A representative area of the concentration maps corresponding to these four species is shown in Figure 2.3. Again, these images show the ability of multivariate data analysis to generate autofluorescence-free images of live cells and interrogate multicolor fluorescence protein species with a

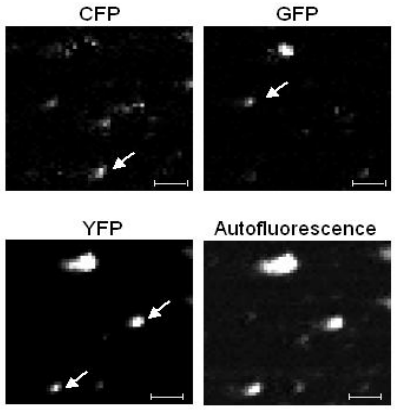


**Figure 2.3. Individual component concentration maps generated from multivariate analysis of hyperspectral images of live yeast cells expressing multicolor**

single laser scan. Cells expressing one color can be located as well as clusters of cells expressing two or three colors reinforcing the need for higher spatial resolution to facilitate single-cell detection. As in the FISH example, future experiments could easily include more colors of fluorescent proteins and/or organic fluorophores provided they could be excited with a single laser.

## 2.3. CONCLUSIONS AND FUTURE WORK

These results demonstrate the power of hyperspectral imaging and multivariate data analysis to understand sources of interfering emissions and increase throughput and reliability of image data. In the case of our microarray imaging we have used this information in a quality control mode to minimize extraneous contaminant emissions at the time they are produced. Preliminary examinations of cell- based systems have been exploratory in nature and more work is needed to determine the limitations of the technique (maximum number of fluorophores, etc) and quantitate the advantages over filter microscopy.

In all examples illustrated here the computational time for the analysis was less than the image acquisition time. However, a potential disadvantage of hyperspectral imaging is large amount of data collected. A hyperspectral scan of an entire microarray at 10 μm resolution could easily be several gigabytes in size. As a result we are continually developing data compression techniques and more efficient algorithms to perform hyperspectral analysis of the large area scans at the diffraction limited spatial resolution our recently upgraded hyperspectral imaging system will provide. With these future improvements hyperspectral imaging will be poised to tackle challenging bioscience microscopy applications.

## 2.4. ACKNOWLEDGEMENTS

The authors acknowledge Gary Jones for his invaluable assistance in constructing the scanner. Sandia is a multiprogram laboratory operated by Sandia Corporation, a

Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-4AL85000.

## 2.5. REFERENCES

[1] RA Schultz, T Nielsen, JR Zavaleta, R Ruch, R Wyatt, and H Garner, "Hyperspectral imaging: A novel approach for microscopic analysis," *Cytometry*, vol. 43, pp. 239-247, 2001.

[2] ME Dickinson, GH Bearman, S Tille, R Lansford, and SE Fraser, "Multi-spectral imaging and linear unmixing add a whole new dimension to laser scanning fluorescence microscopy," *Biotechniques*, vol. 31, pp. 1272-1278, 2001.

[3] MB Sinclair, JA Timlin, DM Haaland, and M Werner-Washburne, "Design, construction, characterization, and application of a hyperspectral microarray scanner," *Applied Optics*, in press, 2003.

[4] PG Kotula, MR Keenan, and JR Michael, "Automated analysis of SEM X-Ray spectral images: a powerful new microanalysis tool.," *Microscopy & Microanalysis*, vol. 9, pp. 1-17, 2003.

[5] DM Haaland, JA Timlin, MB Sinclair, MH Van Benthem, MJ Martinez, AD Aragon, and M Werner-Washburne, Multivariate curve resolution for hyperspectral image analysis: applications to microarray technology," presented at Spectral Imaging: Instrumentation, Applications, and Analysis, San Jose, CA, 2003.

[6] MK Kerr and GA Churchill, "Experimental design for gene expression microarrays," *Biostatistics*, vol. 2, pp. 183-201, 2002.

[7] WJ Wang, S Ghosh, and SW Guo, "Quantitative quality control in microarray image processing and data acquisition," *Nucleic Acids Research*, vol. 29, pp. U32-U39, 2001.

[8] MJ Martinez, AD Aragon, AL Rodriguez, JM Weber, JA Timlin, MB Sinclair, DM Haaland, and M Werner-Washburne, "Identification and removal of contaminating fluorescence from commercial and in-house printed DNA microarrays," *Nucleic Acids Research*, vol. 31, pp. e18, 2003.

[9] KP Doyle, RP Simon, A Snyder, and MP Stenzel-Poore, "Working with GFP in the brain," *Biotechniques*, vol. 34, pp. 492-494, 2003.

[10] YA Labas, NG Gurskaya, YG Yanushevich, AF Fradkov, KA Lukyanov, and MV Matz, "Diversity and evolution of the green fluorescent protein family," *Proceedings of the National Academy of Sciences*, vol. 99, pp. 4256-4261, 2002.

# Chapter 3

# Inferring Genetic Network from Microarray Data

This work was done in collaboration with Margaret Werner-Washburne, University of New Mexico, Department of Biology, Albuquerque, NM, 87131.

**Abstract**

*In theory, it should be possible to infer realistic genetic networks from time series microarray data. In practice, however, network discovery has proved problematic. The three major challenges are 1) inferring the network; 2) estimating the stability of the inferred network; and 3) making the network visually accessible to the user. Here we describe a method, tested on publicly available time series microarray data, which addresses these concerns.*

## 3.1. INTRODUCTION

The inference of genetic networks from genome-wide experimental data is an important biological problem which has received much attention. Approaches to this problem have typically included application of clustering algorithms [6]; the use of Boolean networks [12, 1, 10]; the use of Bayesian networks [8, 11]; and the use of continuous models [21, 14, 19]. Overviews of the problem and general approaches to network inference can be found in [4, 3].

Our approach to network inference is similar to earlier methods in that we use both clustering and Boolean network inference. However, we have attempted to extend the process to better serve the end-user, the biologist. In particular, we have incorporated a system to assess the reliability of our network, and we have developed tools which allow interactive visualization of the proposed network.

## 3.2. NETWORK INFERENCE

The first step in our inference algorithm involves clustering the time series microarray data. The clustering algorithm uses force directed graph layout, and produces a two-dimensional representation of the genes from the microarray [2, 13]. In this representation, genes with similar expression profiles are placed near each other, and genes with different expression profiles are placed farther apart. We then partition this representation using the well-known *k*-means algorithm to provide *k* groups of co-regulated genes. This process not only simplifies the task of network inference (by reducing the problem size), but also results in a network of gene groups, instead of actual genes. These gene groups, which we call *meta-genes*, make the biological analysis and interpretation of the inferred network tractable. Figure 3.1 illustrates the process of obtaining the gene groups.

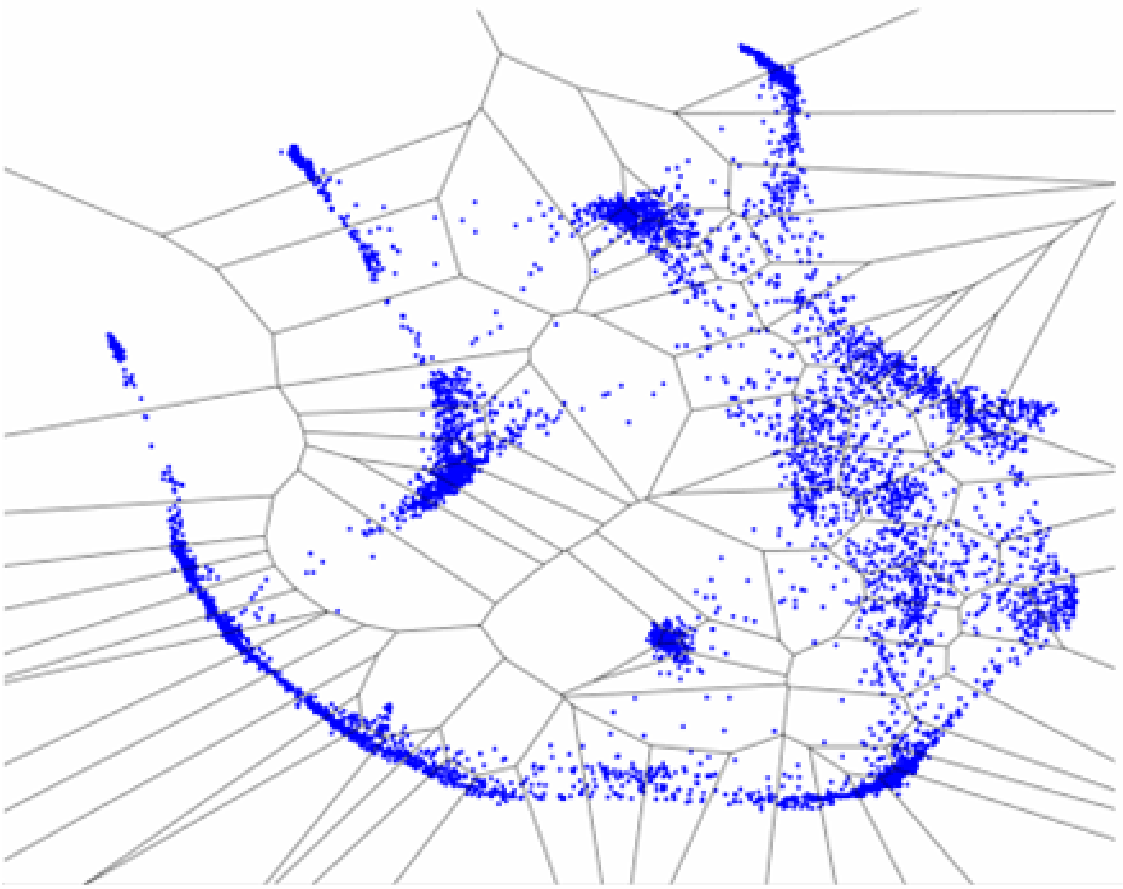

Figure 3.1: Gene map partitioned by *k*-means for yeast time series microarray data in [18].

Since our network inference algorithm is Boolean, we must first discretize the expression levels of our meta-genes. This discretization is accomplished in two steps. First, Support Vector Regression [17] is used to obtain a single continuous curve representing each meta-gene. Next, an on/off expression profile is obtained by thesholding the resulting continuous curve, as shown in Figure 3.2.

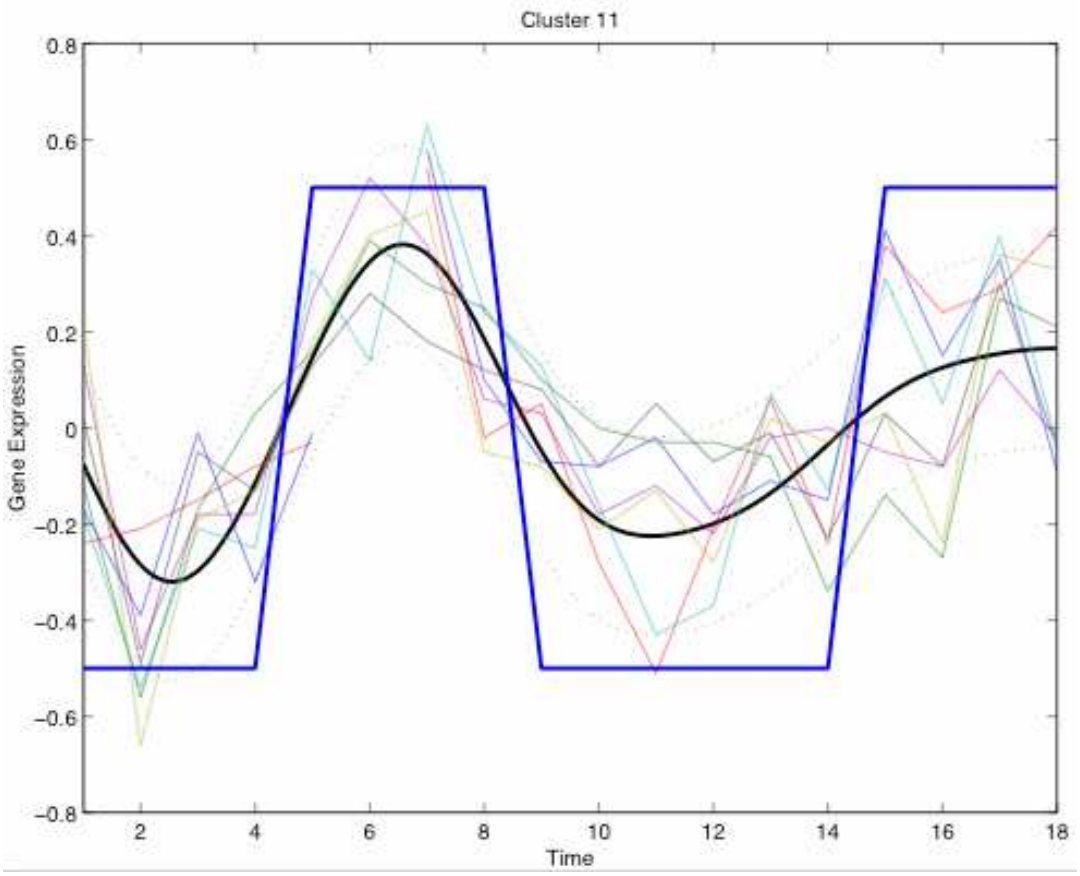


Figure 3.2: Discretized meta-gene for a gene group from Figure 3.1.

After discretizing the meta-genes, we infer a Boolean network. The inference algorithm is based on previous work in chemical reaction network generation [7] and contains routines to count, enumerate, and sample Boolean networks matching the clustered and discretized expression profiles. The inference routines run in $O(2^k n^{k+1})$ time, where $n$ is the number of meta-genes available, and $k$ is the maximum connectivity of a given gene.

In order to more easily interpret the results of our Boolean network inference algorithm, we exploit available tools for electronic circuit analysis. In particular, we perform a two-level Boolean minimization on the truth table representation of the inferred gene network using *Espresso*, a well-known logic simplification tool available from www-cad.eecs.Berkeley.edu. *Espresso* produces a minimized truth table for each meta-gene. Since each meta-gene is processed in the same manner, we get a minimized representation

of the entire network. This new version of the network simplifies the biological analysis and interpretation.

### 3.3. STABILITY ASSESSMENT

Even though the number of possible logic clauses per meta-gene is limited, a large number of possible networks can be inferred from the same meta-genes. To explore the distribution of possible networks, we expand our logic clause calculation to a set of 1000 randomly sampled networks. We use this calculation to generate statistics which identify the most reliable meta-genes and associated clauses.

We also cluster the sampled networks according to their dynamics. Briefly, we cluster two networks when one network differs from another by a pre-defined hamming distance, as measured using its dynamic expression profile. In other words, two networks having different topologies are clustered if they have similar dynamics. Tests on random networks with different sizes and hamming distance thresholds indicate that for a number of unclustered networks (ranging between 1 and $3^{10}$ nodes), the number of clusters is no greater than 500.

Finally, we simulate our inferred network using a continuous model called *BioXyce*, which is a parallel electric circuit simulation tool adapted to biological problems [15]. Results are comparable to the original discretized signal. We note that the simulation was not possible using traditional CMOS-based Boolean logic, but we found that a non-CMOS based logic was successful [16].

## 3.4. NETWORK VISUALIZATION

After the network has been inferred, converted into a minimal set of logical clauses, and been assessed for quality, we present the results in a format amenable to interactive viewing. First, we draw the network using the *dot* graph drawing tool [9], as shown in Figure 3.3. This tool was programmed to use various colors and shapes to encode information specific to the particular application.

To make the drawing interactive, we display it using a web-browser, where each meta-gene is hot-linked and has mouse-over capability. In particular, clicking a meta-gene opens a spreadsheet containing the annotation for the genes in that group, and when a meta-gene is under the mouse, a window pops up to show the original gene expression patterns and corresponding discretization, as shown in Figure 3.2.

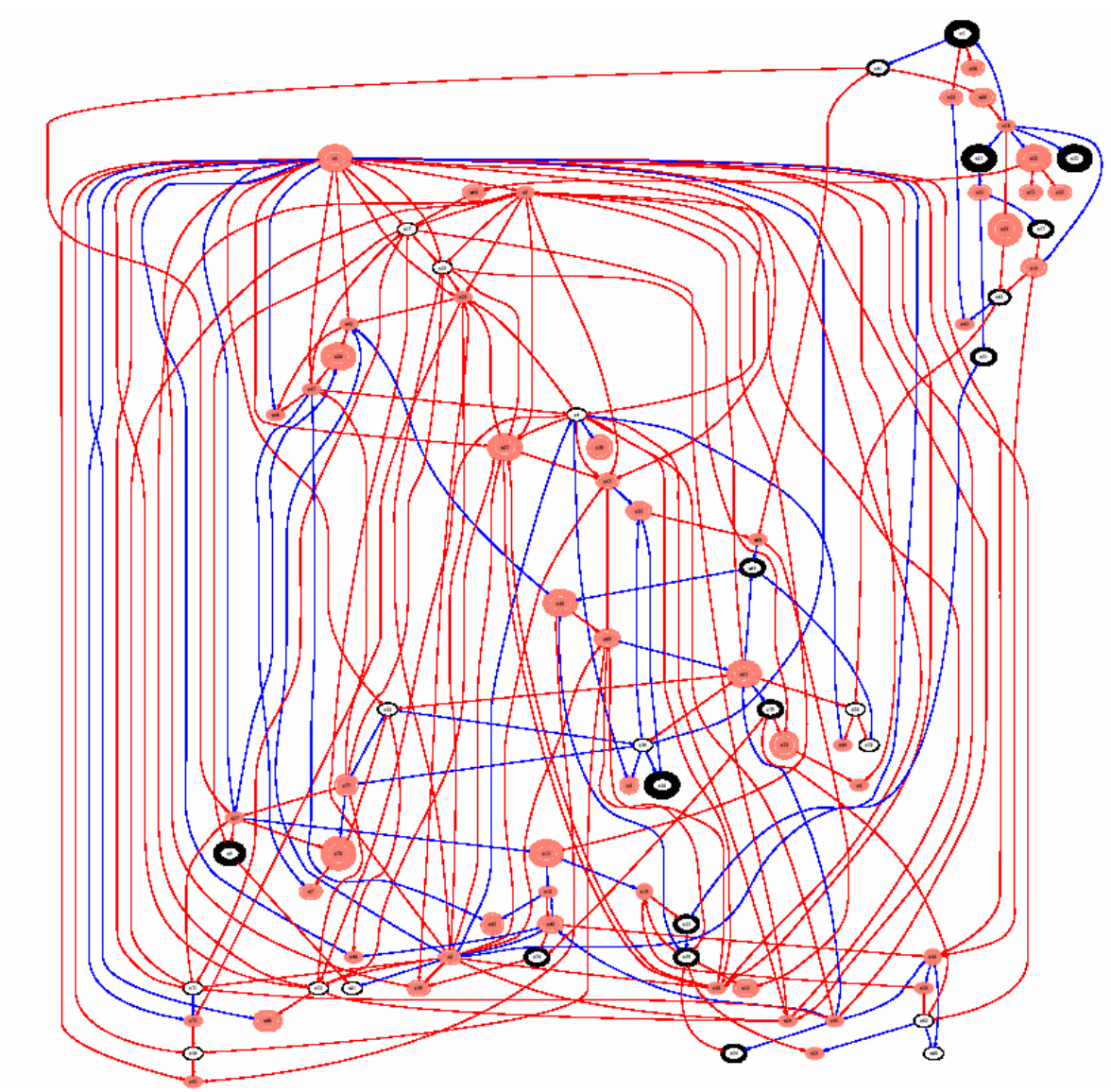

Figure 3.3: Visualization of final network using the yeast time series data from [18].

## 3.5. RESULTS

We have applied our method to the publicly available yeast time series microarray data in [18]. The steps in the process have been illustrated in Figures (3.3.1-3.3).

In Figure 3.1., we used the clustering of the time series data previously performed in [20], along with the partitioning by *k*-means. In this case, we used $k$=100, and discarded clusters with fewer than 20 genes, leaving 81 meta-genes.

In Figure 3.2, we used Support Vector Regression with a Gaussian kernel ($\gamma$=2) and an [epsilon]-tube width of one and a half times the average standard deviation of the expression values at each time point.

In Figure 3.3, we used different color lines for inhibition and activation connections, and different color nodes for essential genes. We used circular nodes for genes involved in the cell-cycle, oval nodes for gene not involved in the cell-cycle, and circles around a node to indicate confidence in the relationships for that node. We computed the confidence bands for a given meta-gene in the network using the cumulative distribution of logical clauses from 1000 networks. We found that 14% of the activation/inhibition clauses appeared in all networks, while 45% of the clauses were present in half of the networks. This result indicates that even while a large number of networks can be inferred, there is some consistency across networks.

Finally, the real proof that our method is useful must come from the analysis and interpretation of the final network. Working with our biological collaborators, we have developed two testable hypothesis based on our proposed network. First, we discovered that the meta-gene module in the upper right corner of Figure 3.3 consists almost entirely of genes involved in exit from alpha-arrest. These cells were exposed to alpha mating factor, which stops the cell-cycle at a checkpoint until it is removed, thereby providing a way to synchronize the cells in the growth medium. The gene groups in the upper right of the drawing seem to be involved in this synchronization process.

Second, we noticed that many of the links in the drawing are inhibitory. This unexpectedly large number of inhibitory controls goes counter to the currently accepted regulatory

model and may suggest that genetic networks are more tightly controlled than has been previously assumed. Further experiments, both laboratory and computational, will be necessary to test these hypotheses.

## 3.6. FUTURE WORK

We have two primary objectives for the immediate future. First, we have already starting analyzing the stability of our methods in greater detail. In particular, the circles around the nodes in Figure 3.3 are meant to give an indication of likelihood that a given meta-gene will have the same relationships to other meta-genes in alternate networks generated by the network inference algorithm. We plan to make these computations much more robust by using bootstrapping methods [5] to assess the variance caused by changes in our sampling algorithms. These changes include altering the curve-fitting and discretization parameters as well as considering even more alternate inferences provided by the network inference algorithm.

Second, we intend to perform a full and thorough analysis of time series microarray data that has been collected by a collaborator (A. Martino) in order to infer T-cell regulatory networks. In particular, we will study T-cell regulatory networks triggered through tyrosine kinase receptor activation.

## 3.7. CONCLUSIONS

The development of this network and visualization environment has required the collaboration of researchers in math (JLF, SM), computer sciences (GD, EM), and yeast genomics (MWW). From the beginning we have focused on *the entire network inference process*. We have developed clustering, discretization, and inference algorithms, and have attempted to validate their output. Finally, we have presented the results using an interactive network browser for accessible biological interpretation. Although we will

continue to improve our process, it has already yielded two testable biological hypotheses, one concerning exit from arrested states, and one concerning the level of control present in genetic networks.

## 3.8. ACKNOWLEDGEMENTS

This work was funded by Sandia Laboratory Directed Research and Development project 52533. Some of the related work was funded by the US Department of Energy's Genomics: GTL program (www.doegenomestolife.org) under project, "Carbon Sequestration in *Synechococcus Sp.*: From Molecular Machines to Hierarchical Modeling," (www.genomes-to-life.org). Sandia is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

## 3.9. REFERENCES

[1] T. Akutsu, S. Miyano, and S. Kuhara. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. In *Pacific Symposium on Biocomputing*, volume 4, pages 17–28, 1999.

[2] G. Davidson, B. N. Wylie, and K. W. Boyack. Cluster stability and the use of noise in interpretation of clustering. In *IEEE Symposium on Information Visualization (INFOVIS)*, 2001.

[3] H. de Jong. Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.*, 9(1):67–103, 2002.

[4] P. D'haeseleer, S. Liang, and R. Somogyi. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16(8):707–726, 2000.

[5] B. Efron. Bootstrap methods: Another look at the jackknife. *Ann. Statistics*, 7(37):1–26, 1979.

[6] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression data. *Proc. Nat. Acad. Sci.*, 95(25):14863–14868, 1998.

[7] J.-L. Faulon and A. G. Sault. Stochastic generator of chemical structure. 3. reaction network generation. *J. Chem. Inf. Comput. Sci.*, 41(4):894–908, 2001.

[8] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *J. Comput. Biol.*, 7:601–620, 2000.

[9] E. R. Gansner, E. Koutsofios, S. C. North, and K. P. Vo. A technique for drawing directed graphs. *IEEE Trans. on Soft. Eng.*, 19(3):214–230, 1993.

[10] J. Goutsias and S. Kim. A nonlinear discrete dynamical model for transcriptional regulation: Construction and properties. *Biophys. J.*, 86:1922–1945, 2004.

[11] D. Husmeier. Reverse engineering of genetic networks with bayesian networks. *Biochem. Soc. Trans.*, 31:1516–1518, 2003.

[12] S. A. Kauffman. *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, New York, 1993.

[13] S. Kim et al. A gene expression map for *c. elegans*. *Science*, 293:2087–2093, 2001.

[14] J. C. Liao et al. Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Nat. Acad. Sci.*, 100(26):15522–15527, 2003.

[15] E. E. May and R. Schiek. Simulating regulatory networks using *xyce*. In *Fourth International Conference on Systems Biology*, 2003.

[16] H. H. McAdams and A. Arkin. Simulation of prokaryotic genetic circuits. *Annual Review of Biophysics and Biomolecular Structure*, 27:199–224, 1998.

[17] A. J. Smola and B. Scholkopf. A tutorial on support vector regression. NeuroCOLT Technical Report NC-TR-98-030, Royal Holloway College, University of London, UK, 1998.

[18] P. T. Spellman et al. Comprehensive identication of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by micraoarray hybridization. *Molecular Biology of the Cell*, 9:3273–3297, 1998.

[19] J. Tegner, M. K. Yeung, J. Hasty, and J. Collins. Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling. *Proc. Nat. Acad. Sci.*, 100(10):5944–5949, 2003.

[20] M. Werner-Washburne et al. Concurrent analysis of multiple genome-scale datasets. *Genome Research*, 2002.

[21] M. K. Yeung, J. Tegner, and J. Collins. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Nat. Acad. Sci.*, 99(9):6163–6168, 2002.

# Chapter 4

# Dynamical Robustness in Gene Regulatory Networks


## ABSTRACT

*We investigate the robustness of biological networks, emphasizing gene regulatory networks. We define the robustness of a dynamical network as the magnitude of perturbation in terms of rates and concentrations that will not change the steady state dynamics of the network. We find the number of dynamical networks versus their dynamical robustness follows a power law. We observe module robustness to increase with node degree in published gene regulatory networks. Finally, based on dynamical robustness, we propose a growth model for producing networks with power law degree distributions.*


## 4.1. INTRODUCTION

Many biological, social and technological networks have scale free characteristic, that is, the degree of the nodes follows a power law distribution. This differs from random networks where node degrees follow a Poisson distribution. In order to generate random graphs with power law distributions, specific schema have been developed, such as growth through preferential attachment [1], or node duplication followed by edge deletion [2]. There is still much debate about whether or not these growth models are appropriate when dealing with gene and protein networks. In addition, these models do not take dynamics into account, even though dynamical robustness is an important aspect

of gene and protein networks. Quoting Uri Alon [3], biological networks are robust to component tolerance and this should impose severe constraints on their design. Furthermore, it has been shown that power law networks exhibit robust behavior for power law exponents greater than two [4], and such exponent values are generally observed in most published protein and gene networks. In this paper, we further explore the relationship between dynamical robustness and scale free properties.

## 4.2. ROBUSTNESS

We focus on gene regulatory networks and define the robustness of such networks as the magnitude of perturbation (in rates and concentrations) that can be carried out without changing the steady state dynamics of that network. If the network is limited to the Boolean activation/inhibition model, robustness becomes the number of different networks having the same set of attractors. Indeed, as illustrated in Figure 4.1, different networks can lead to the same steady state dynamics.

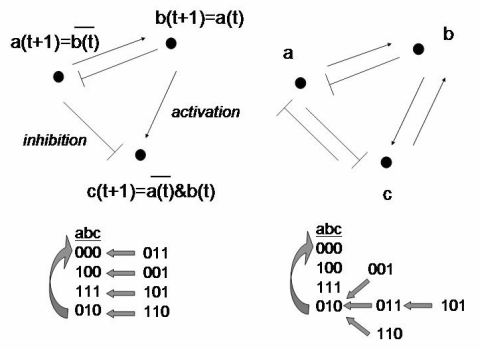


Figure 4.1. Networks with identical attractors. Both networks cycle in the attractor composed of states (000,100,111,010).

## 4.3. RANDOM NETWORK ROBUSTNESS

In order to search for robust networks, we have systematically generated all activation/inhibition networks with up to five nodes. Undirected graphs were first

enumerated using McKay's orderly enumeration algorithm [5]. Next, for each enumerated graph, the two possible edge directions and the activation/inhibition labels were added in all possible ways. All resulting non-isomorphic networks were then run using Boolean dynamics. Each run was performed over all possible initial conditions until an attractor was reached. A binary string was compiled representing the Boolean states of the nodes within the attractors. That string was canonized considering all node permutations. Finally, networks were clustered according to their dynamical canonical string. Figure 4.2 plots the number of clusters versus cluster sizes.

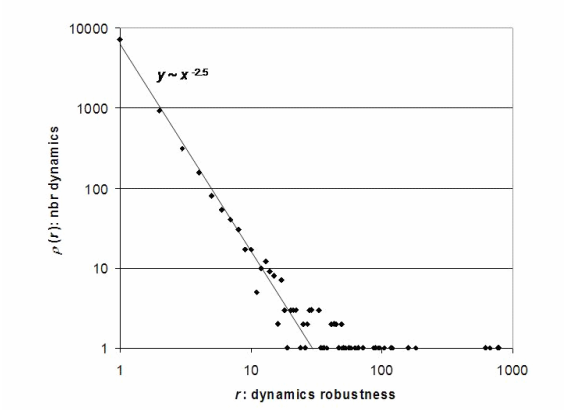


Figure 4.2. Number of dynamics vs. dynamical robustness for activation/inhibition networks.

Figure 4.2 demonstrates that most networks do not share their steady state dynamics with others, while a few networks are highly robust. Precisely, the number of different steady state dynamics versus the dynamic robustness seems to follow a power law with exponent 2.5. This result is surprising, since according to the theory of random graphs, network characteristics should follow Poisson distributions. We hypothesize that power law behavior observed in biological networks is a consequence of dynamic robustness.

## 4.4. GENE NETWORK ROBUSTNESS

To further test our hypothesis we analyzed the dynamics of three activation inhibition networks: a transcriptional regulation network for E-coli [6], and two Yeast gene

regulatory networks ([7], and a network we inferred from microarray data [8]). In all of these networks, we found that the number of subgraphs (or modules) with up to five nodes increased with the dynamical robustness.

These results can be explained by considering that biological networks are composed of modules connected together [9], and that networks composed of modules can be constructed with a power law degree distribution, $P(k)$, if the modules have a fitness (robustness in the present case) also following a power law, $\rho(r)$. Precisely, Caldarelli *et al.* [10] have shown that networks of $N$ modules can be constructed with the following distribution of node degrees $P(k) = r_M^2/(N\langle r\rangle)\rho[r_M^2/(N\langle r\rangle)\ \ k]$, where $\langle r\rangle$ and $r_M^2$ are the averaged and maximum robustness of the modules. Note that $P()$ follows a power law as long as $\rho()$ does. The growth of such networks is simply carried out by linking modules, one with robustness $r$ and the other with robustness $s$, with probability $rs/r_M^2$. Figure 4.3 was compiled for a network of 150,000 modules grown using the above probability and following the robustness distribution of Figure 4.2. Clearly both figures follow the same power law distribution. Thus, we conclude that the distribution of module robustness can be used to explain the node degree distribution found in gene regulatory networks.

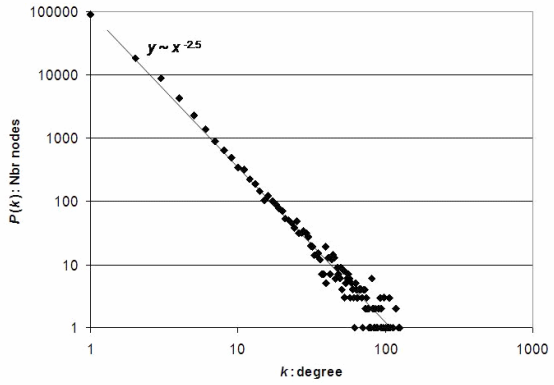


Figure 4.3. Degree distribution for networks generated using $\rho(r) \sim r^{-2.5}$ and edge probability $rs/r_M^2$.

## 4.5. ACKNOWLEDGEMENT

The authors are pleased to acknowledge funding provided by the US Department of Energy's GTL program under project, "Carbon Sequestration in *Synechococcus Sp*.: From Molecular Machines to Hierarchical Modeling" (http://www.genomes2life.org/).

## 4.6. REFERENCES

[1] A.-L. Barabasi, R. Albert, "Emergence of scaling in random networks", *Science*, 286,1999, 286, pp. 509-512.

[2] S. A. Teichmann, M. M. Babu, "Gene regulatory network growth by duplication", *Nature Genetics*, 36, 2004, pp. 492-496

[3] U. Alon, "Biological Networks: The Tinkerer as an Engineer", *Science*, 301, 2003, pp. 1866-1867.

[4] M. Aldana, P. A. Cluzel, "Natural class of robust networks", *PNAS,* 100, 2003, pp. 8710-8714

[5] B. D. McKay, "Isomorph-Free Exhaustive Generation*", J. of Algorithms*, 13, 1992, pp. 306-320.

[6] S. S. Shen-Orr et al. "Network motifs in the transcriptional regulation network of E-coli", *Nature Genetics*, 31, 2002, pp. 64-68.

[7] N. Guelzim *et al*. "Topological and causal structure of the yeast transcriptional regulatory network", *Nature Genetics*, 31, 2002, pp. 60-63.

[8] S. Martin *et al*. "Inferring Genetic Networks from Microarray Data", *in bib.*

[9] A. Vespignani, "Evolution thinks modular", *Nature Genetics*, 35, 2003, pp. 118-119.

[10] G. Caldarelli, *et al*., "Scale-Free Networks from Varying Vertex Intrinsic Fitness," Phys. Rev. Lett., 2002.

# Chapter 5

# A Genome-Wide Transcriptional Analysis of Murine T Cells in Response to IL-2 Stimulation


**ABSTRACT**

*Using an Affimetrix GeneChip mouse genome, we collected microarray data at 12 series time points, for a genome-wide analysis of Murine T Cells in Response to IL-2 stimulation. The first time point was collected after T cells had been cultured without IL-2 for 14 hours. Immediately after the first time point, IL-2 was added and data were collected for IL-2 stimulated T cell at 15 minutes, 30 minutes, 1 hour, every 2 hours from 2 to 12 hours, 16 hours, and 24 hours. We clustered the gene expression profiles over the 12 times points into 23 meta-genes, these 23 meta-genes were found to belong to four categories, early genes (up-regulated after 1 hour), cyclic genes (up-regulated after 1 hour, then down-regulated, and up-regulated again after 16 hours), intermediate gene (up-regulated after 2 hours), and late genes (up-regulated after 8 hours). From the clustered expression profiles we inferred a network of activation and inhibition relationships between the meta-genes. This network was analyzed using the Gene Ontology (GO) annotations of the mouse genome. Our microarray data and corresponding inferred network indicate that after IL-2 stimulation, T cell undergo a cell cycle leading to cell division and proliferation.*

## 5.1. INTRODUCTION

A major feature of the immune response to bio-toxin (e.g., a foreign antigen) or pathogen is the rapid proliferation of T cells. The initial signal induced by a foreign antigen is the production of interleukin 2 (IL-2) upon recognition of the antigen by the T cell receptors (TCR). IL-2 promotes T cell proliferation by activating cell growth pathways and inducing promitogenic genes such as c-myc, cyclin D2 and cyclin E (Nelson and Willerford 1998). IL-2 is a four α helix cytokine produced mainly by activated $CD4^+$ T cells (Letterio and Roberts, 1998). IL-2R is a surface complex composed of three proteins: IL-2Rα, IL-2Rβ and $\gamma_c$ (Takeshita et al. 1992). Upon IL-2 binding to IL-2Rα, the IL-2R activates tyrosine kinase JAK1 and JAK3, which associate with the cytoplasmic domains of the IL-2R β and $\gamma_c$, respectively (Gorelik and Flavell 2002), and initiates IL-2 intracellular signaling pathways. IL-2Rβ undergoes phosphorylation on tyrosine residue $Y^{338}$ and $Y^{510}$, which are docking sites for the adaptor protein Shc and the transcription factor STAT5, respectively. Two IL-2 signal pathways are activated, one is mediated by adaptor Shc, the other is by transcriptional factor STAT5 (Friedmann et al. 1996). Subsequently, Grb2/Sos complex is recruited to Shc and the Ras/mitogen-activated protein kinase pathway is activated, alternatively GAB2 is recruited to Shc and the phosphatidylinositol 3-kinase (PI3K) signaling pathway is activated (Ravichandran et al. 1995, Schaeffer and Weber 1999, Gu et al. 2000). In contrast, STAT5 undergoes tyrosine phosphorylation, followed by dimerization and translocation to the nucleus (Lin and Leonard 2000). Both the Shc and STAT5 signaling pathways can induce promitogenic gene expression and cell proliferation in a redundant manner.

DNA array analysis identified 47 genes in human T cells as IL-2 regulated early genes with 2 fold change at expression level after 4 hours IL-2 stimulation, 19 genes were up-regulated such as IL-4R, pim1 and bcl2, the rest were down-regulated genes such as cyclin-dependent kinase inhibitor gene $P27^{(Kip1)}$ (Beadling and Smith 2002). A number of promitogenic genes such as c-fos, c-jun, c-myc, cyclin D2, cyclin E, bcl-x, bcl2, IL-2Rα, and pim1 are known to be induced by IL-2 in murine T cells (Friedmann et al. 1996, Lord et al. 1998, Moon and Nelson 2001, Lord et al. 2000, Matikainen et al. 1999). Now the

mouse genome sequence has been completed, and Affymetrix Mouse Genome Array covering the mouse genome is available. Global gene expression profile analysis in response to IL-2 in mouse T cells can thus be performed to infer the network of relationships between genes responsible for the complicated cellular process in T cell proliferation. The identified genes responsive to IL-2 could in turn be used as signatures in the immune response to bio-toxin or pathogen. Here we report the global gene expression profile in the IL-2 dependent murine T cells line CTLL-2 in response to IL-2 stimulation from 30 minutes to 24 hours in 12 time points series.

The report is organized as follows, we in section 2, we describe materials and methods including cell culture, RNA isolation, gene expression analysis, clustering, and gene network inference. In section 3 that follows we present our results for the 12 times points series microarray data and in particular we identify known IL-2 regulated genes, we cluster and discretize the microarray data, and we infer networks of activation and inhibition relationships between cluster of genes. The clusters of genes and the inferred networks are discussed in section 4 in using the Gene Ontology (GO) annotations.

## 5.2. MATERIALS AND METHODS

### 5.2.1. Cell culture

The murine T cell line CTLL-2 was obtained from American Type Culture Collection (ATCC, Manassas, VA). Mouse CTLL-2 cells is interleukin (IL-2) dependent cytoxic T lymphocyte and was maintained in ATCC modified RPMI-1640 medium (10 mM HEPES, 2 mM L-glutamine, 1 mM Sodium pyruvate, 4.5 g/L glucose, 1.5 g/L sodium bicarbonate), supplemented with 10% fetal bovine serum, 2 mM L-glutamine, 50 U/mL penicillin, 50 mg/mL streptomycin, 1 mM sodium pyruvate, 45 μM β-mercaptoethanol, and 50 U/mL human rIL-2 (Chiron, Emeryville, CA). Cells were cultured in an incubator with 5% carbon dioxide (CO2) at 37$^{o}$C to mid-log phase (~2 X $10^{5}$ cells/mL), and collected by centrifugation, washed with phosphate buffered saline (PBS) (Invitrogen, Carlsbad, CA), resuspended at ~1 X $10^{6}$ cells/mL in the medium described above without IL-2 (IL-2 starvation) and cultured for 14 hours at 37$^{o}$C with 5% CO2. Cells were then

collected (0 hour, no IL-2 stimulation) immediately before IL-2 was added to 100 U/mL (IL-2 stimulation). Cells were further collected at 11 series time points: 15 minutes, 30 minutes, 1 hour, 2 hours, 4 hours, 6 hours, 8 hours, 10 hours, 12 hours, 16 hours and 24 hours after IL-2 stimulation. Cell pellets were washed with PBS to removed RPMI-1640 medium, resuspended in 80 µL PBS before adding 0.6 mL RNAlater RNA stabilization solution (Qiagen, Valencia, CA), stored at -80$^{o}$C freezer.

### 5.2.2. Total RNA isolation

Total RNA was isolated using RNeasy Mini Kit (Qiagen, Valencia, CA) according to manufacturer's protocol. Briefly, cells was pellet from cells in RNA stabilization reagent, disrupted by vortexing after adding 600 µL buffer RLT, pipet onto QIAshredder spin column to homogenize the sample (12k rpm centrifugation for 2 minutes). 600 µL 70% ethanol was added to the lysate, mixed and pipet onto an RNeasy mini column, centrifuged at 12k rpm for 30 seconds. After washing with 350 µL buffer RW1, RNase-free DNase I (10 µL DNase + 70 µL buffer RDD) was pipet onto RNeasy silica-gel membrane and incubated for 15 minutes at room temperature, washed with 350 µL buffer RW1, followed by two times wash with 500 µL buffer RPE. The total RNA was eluted from the RNeasy column with 30 µL RNase-free dH2O, and stored at -80$^{o}$C freezer.

### 5.2.3. Gene expression analysis

Affymetrix GeneChip Mouse Genome 430 2.0 Array (Affymetrix, Santa Clara, CA) was used for gene expression analysis. The array provides complete coverage of the transcribed mouse genome, 45,000 probe sets to analyze the expression level of over 39,000 transcripts and variants from over 34,000 well characterized mouse genes.

Gene expression analysis was carried out at Stanford PAN Biotechnology Facility (http://cmgm.stanford.edu/pan/gene/index.html). Specifically, total RNA from each time point was used to prepare biotinylated target cRNA according to the manufacturer's manual (http://www.affymetrix.com/support/technical/manual/expression_manual.affx), three replicate experiments were done for each time point. Briefly, ~ 4 µg total RNA was used for the first-strand and second cDNA synthesis using the One-Cycle cDNA

synthesis Kit (Affymetrix, Santa Clara, CA), the first-strand cDNA was synthesized by reverse transcription using a T7-Oligo(dT) primer and SuperScript II, the second-strand cDNA was synthesized using E. coli DNA polymerase I and ligase. Synthesis of biotin-labeled cRNA was performed at 37°C for 16 hours in a thermal cycler using GeneChip IVT Labeling Kit (Affymetrix), biotinylated cRNA generated was further fragmented. Target hybridization was processed following manufacturer's recommendation using the instrument operated by Affymetric GeneChip Operating Software (GCOS) and Microarray Suite version 5.1 (MAS 5.1). Spike transcript controls and fragmented cRNA were added to a hybridization cocktail, 200 μL containing 10 μg fragmented cRNA of the cocktail was hybridized to the expression microarray by incubation at 45°C for 16 hours. Arrays were then washed and stained with streptavidin-phycoerythrin in the Fluidics Station 400 before being scanned on Affymetrix Gene-Chip scanner 3000. After scanning, array images were assessed by eye to confirm scanner alignment and the absence of significant bubbles or scratches on the chip surface. Expression cell intensity data was automatically generated by GCOS. The 3'/5' ratios for GAPDH (glyceraldehyde-3-phosphate dehydrogenase) and beta-actin were confirmed to be within acceptable limits (0.78- 0.97), and BioB spike controls were found to be present on all chips with BioC, BioD, and CreX also present in increasing intensity. Finally, each image was globally scaled to a target intensity of 500 and the scaling factors for all arrays were confirmed to be within acceptable limits (2.2-6.2) as were background and noise.

### 5.2.4. Data analysis

The fluorescent intensity of each probe was quantified using Microarray Suite version 5.1 (MAS 5.1) and Affymetric GeneChip Operating Software (GCOS, version 1.3). The expression level of a single mRNA, defined as the signal, was determined by GCOS, which uses a weighted average fluorescence intensity difference obtained among the 11 to 20 probe pairs that interrogate the expression of each individual gene. This software also makes a detection call (present [P], marginal [M], or absent [A]) for each gene or probe set, based on the consistency of the performance of the individual probe pairs, the hybridization above background, and the signal-to-noise ratio. Two-way comparisons of the microarray data were also performed using GCOS. Specifically, changes in gene

expression between the control cells (time point 0 hour, no IL-2 stimulation) and IL-2 stimulated cells were evaluated at each time point. These comparisons in GCOS provided additional data including the signal log ratio (fold change presented in logarithmic form) and the "change call" (increased [I], decreased [D], marginally increased [MI], marginally decreased [MD], or no change [NC]) for each gene being interrogated. The data were then imported into a Microsoft Excel (Redmond, WA) spreadsheet. To identify genes that exhibited differences in expression between the control cells and IL-2 stimulated cells, the data sets were trimmed in Excel using the following inclusion criteria. For a probe set to be included in this trimmed data set it had to display in all the three replicates: (1) a change call other than no change (NC), (2) the same trend of change call (I, increase, D, decrease), (3) a present call (P) and/or signal intensity ≥100, and (4) at least a 1.5-fold difference in expression between the 2 compared conditions. Additional annotation data were incorporated into the data set using the Affymetrix web-based analysis tool NetAffx.

### 5.2.5. Clustering

The first step to infer relationships between genes involves clustering the time series microarray data. Clustering the microarray data serves two purposes: first, this process simplifies the task of network inference by reducing the problem size; second, the clustering results in a network of gene groups, instead of actual genes. The gene groups, which we call *meta-genes*, make the biological analysis and interpretation of the inferred network more tractable.

We use *k*-means (Jain *et al*., 1999) to cluster our dataset. To decide how many clusters should be produced (the value of $k$), we developed a measure of *internal consistency*. Our measure is defined for a given partition of the dataset using the singular value decomposition (SVD) (Trefethen & Bau, 1997). To define internal consistency, suppose we are given $k$ and we compute a $k$-means partition of our $m \times n$ dataset $X$, where $m$ is the number of time points, and $n$ is the number of genes. For the $j$th cluster ($j = 1,\ldots,k$) we have a matrix $X_j$ of microarray measurements, where the rows are time-points and the columns are genes, so that $X_j$ is a $m \times g_j$ matrix, where $g_j$ is the number of gene in the $j$th

cluster. Using the SVD, we decompose $X_j = U_j S_j V_j^T$, where $U_j$ and $V_j$ are orthogonal matrices and $S_j$ is a diagonal matrix whose entries describe the importance of the columns of $U_j$ and $V_j$.

To be precise, the matrix $S_j V_j^T = U_j^T X_j$ contains the projections of the columns (time courses) of $X_j$ onto the basis $U_j$. The entries of $S_j$ (singular values) give the relative importance of the columns of $U_j$. If the first entry of $S_j$ is much larger than the second entry then we know that most of the information in the columns of $X_j$ is captured by a single dimension. We thus define the internal consistency of the *j*th cluster to be the ratio of the first and second singular values in $S_j$. This is a measure of the correlation between all of the time courses in the *j*th cluster. The internal consistency also provide a measure on how well a single dimension can describe all the time courses. For the problem of network inference we want each of our clusters to have a high internal consistency. We can decide how many clusters we should use by comparing the average internal consistency of different partitions of the dataset (different numbers of clusters) and choosing the clustering with the best average internal consistency.

### 5.2.6. Discretization

Given an appropriate set of meta-genes, we now discretize the meta-gene expression levels. Such a discretization is necessary because we use a Boolean network inference algorithm. Our discretization is accomplished in two steps. First, support vector regression (SVR) (Smola & Scholkopf, 1998) is used to provide a continuous, smooth representation of the genes in a given group. This type of regression is performed by solving a quadratic programming problem and has two parameters: and ε width and a kernel function. The ε width is used to encapsulate the curves in a given meta-gene group within an ε-tube, and the kernel function is used to fit different types of curves (e.g. linear or non-linear). The end result of SVR encapsulates the time courses in a meta-gene group within an ε-tube centered around a smooth curve, where the curve is a linear combination of kernel functions. For this exercise we used Gaussian kernels with a width $\sigma = 1$, and we chose ε to be one and a half times the average standard deviation of the values at each time point.

The second step in our discretization consists of thresholding the curve obtained by the support vector regression. Assuming that the meta-gene group is well represented by the SVR curve, we can produce a discrete version of the meta-gene by thresholding the curve against it's average value: a higher than average meta-gene expression is given a value of 1 (up-regulated), while a lower than average meta-gene expression value is given a value of 0 (down-regulated).

### 5.2.7. Clustering and discretization stability

Going from a full set of microarray to a reduced set of discrete meta-genes is sure to involve some loss of information and/or introduction of error. In our case, the most likely source of error is the clustering step. Not only must we choose an appropriate value for $k$, but we must also examine the effect of different random starting conditions for $k$-means. (It is worth noting the difficulty with the clustering step is not inherent to $k$-means itself, but to clustering methods in general: we would be unlikely to see improvements with a different clustering method.) To examine the effect of random starting conditions, we repeat our entire procedure using a range of values for $k$ as well as a number of random starting conditions for each $k$. We then discretize the resulting meta-genes using SVR as described above for each $k$ and each random starting condition.

We compare the different discretizations using a simple measure of set intersection. We consider a discretization to be a set of discrete time courses, where each time course is a vector, so that a discretization is a set of vectors. If we have two such sets $A$ and $B$, then we can compute their similarity by computing

$$\frac{|A \cap B|}{\sqrt{|A||B|}},$$

where $|\bullet|$ denotes the cardinality of a set. We note that this measure is between 0 and 1 (inclusive), and is 1 if and only if $A = B$.

### 5.2.8. Inference algorithms

In this section we describe the inference algorithms use to infer networks from gene expression profiles. These algorithms count, sample or enumerate all the possible networks matching a given set of discretised profiles. Precisely, the expression profiles must be given for every node in a Boolean form, 0 for down-regulated and 1 for up-regulated. Networks are counted, sampled and enumerated at the node (gene or cluster of genes) level, that is, for every node $v$, one determines all the possible sets of nodes that control the expression profile of $v$. Expression profiles are given over time course and the algorithms can accept several time courses corresponding to different initial conditions. These initial conditions can be different stimuli, or various knockouts experiments.

The algorithms use as input a set of $n$ nodes $V = \{v_1, v_2, \ldots, v_n\}$ which correspond to the meta-genes or clusters of genes determined in a previous step (cf. 2.5), and a set PROFILES, which is composed of several expression profiles corresponding to different initial conditions. For any given profile in the set PROFILES, the expression of every node is given over a specified time course, different profiles do not necessarily have the same time course length. To limit redundant networks to be constructed, different nodes should have different profile but this requirement is not a necessity to run the algorithms.

The basic step used by the algorithms is `INFER-FUNCTION`, that routine determines if a set of nodes $v_1, v_2, \ldots, v_k$ with $k \leq n$ can explain the expression profile of a given node $v$. In other words that routine return the Boolean function by which $v_1, v_2, \ldots, v_k$ control the expression of $v$. This function is empty if $v_1, v_2, \ldots, v_k$ do not control $v$.

```
INFER-FUNCTION(PROFILES, v, v1, …, vk)
f = **…* (string of 2^k characters)
For all profile in PROFILES do
  For all time t in profile do
    input = profile(v1, t) profile(v2, t)…profile(vk, t)
    (note that input is an integer in binary format)
```

```
    if f(input) = *
       then f(input) = profile(v,t+1)
       else if f(input) ≠ profile(v,t+1) return(∅)
    fi
  Done
Done
return(f)
```

As an example of INFER-FUNCTION suppose we are given a six time points course series for three genes $v_1$,$v_2$, and $v_3$. As before, we write 1 when the gene is up-regulated and 0 when down-regulated. The expression of the three genes $v_1$,$v_2$, and $v_3$ are t = 1h: 011, t = 2h: 000, t = 3h: 100, t = 4h: 111, t =5h: 010, and t = 6h: 000. We are interested deriving the function for $v_3$, and we assume that $v_1$ and $v_2$ are potential inputs. Can we explain $v_3$ expression profile using $v_1$ as the only input? The following table can easily be drawn from the expression profiles of $v_1$,$v_2$, and $v_3$:

| t | $v_1$ | t | $v_3$ |
|---|---|---|---|
| 1 | 0 | 2 | 0 |
| 2 | 0 | 3 | 0 |
| 3 | 1 | 4 | 1 |
| 4 | 1 | 5 | 0 |
| 5 | 0 | 6 | 0 |
| 6 | 0 | | |

When $v_1$ is down-regulated $v_3$ is also down-regulated at the next time step, however when $v_1$ is up-regulated $v_3$ can be both up- and down-regulated. Thus $v_1$ cannot explain $v_3$. The same is true if $v_2$ is used as input for $v_3$. Now let's us examine that case when $v_1$ and $v_2$ are both input of $v_3$. The following table is drawn from the expression profiles:

| t | $v_1$ | $v_2$ | t | $v_3$ |
|---|---|---|---|---|
| 1 | 0 | 1 | 2 | 0 |
| 2 | 0 | 0 | 3 | 0 |
| 3 | 1 | 0 | 4 | 1 |
| 4 | 1 | 1 | 5 | 0 |

| | | | | |
|---|---|---|---|---|
| 5 | 0 | 1 | 6 | 0 |
| 6 | 0 | 0 | | |

In the above case one notes that when $v_2$ is up-regulated $v_3$ is down-regulated the next time step and this disregarding $v_1$. Now, when $v_2$ is down-regulated, $v_3$ is up-regulated if $v_1$ was up-regulated at the previous time step, and $v_3$ is down-regulated if $v_1$ was. This indicates that $v_2$ inhibits $v_3$ and $v_1$ activates $v_3$. The Boolean function controlling $v_3$ corresponding to the input $v_1v_2$ = 00, 01, 10, 11 is f($v_3$) = 0010, written in a different form f($v_3$) = $v_1$ AND NOT $v_2$.

Using INFER-FUNCTION, inferring network can easily be done by processing each node one after another. The pseudo code given below uses three inputs, the provided expression profiles, the corresponding set of nodes *V*, and a parsimonious flag. When the parsimonious flag is turned on the algorithm produces for each node only the minimum number of connections. The possible connections and associated Boolean functions are stored for each node *v*, in a set INPUTSET(*v*). That set is thus composed of a series of k+1-tuples, each comprising the list of k inputs and as the k+1 entry the associated Boolean function (returned by INFER-FUNCTION).

```
INFER-NETWORK (PROFILES, V, parsimonious)
For all nodes v of G do
   INPUTSET(v) = ∅
   For k = 1 to |V| do
     For all k-tuples v1,…,vk do
         f = INFER-FUNCTION(PROFILES,v,v1, …,vk)
         if f ≠ ∅ then
            add (v1,…, vk , f) to INPUTSET(v)
         fi
     done
     if parsimonious and at least one function f
          was found for node v then
        k = |V| + 1
```

```
      fi
   done
done
return(INPUTSET)
```

To count the number of possible networks matching a given set of expression profiles one first run INFER-NETWORK and then run the following algorithm:

```
INPUTSET = INFER-NETWORK(PROFILES,V,parsimonious)
COUNT-NETWORK (V,INPUTSET)
P = 1
For all nodes v of V do
     P = P . |INPUTSET(v)|
done
```

Note that the count is simply the product of the number of possible inputs for each node.

To sample network one runs first INFER-NETWORK and then select at random for each node one of its possible input.

```
INPUTSET = INFER-NETWORK(PROFILES,V,parsimonious)
SAMPLE-NETWORK (V,INPUTSET)
For all nodes v of V do
     select at random an element input of INPUTSET(v)
     print v, input
done
```

Finally, to enumerate all networks, one runs INFER-NETWORK and lists and prints all possible inputs for each node. Recall that INFER-NETWORK generates for each node a

list of possible inputs and associated Boolean functions. In order to enumerate all possible networks one enumerates all possible inputs one after another, this necessitates maintaining a variable for each node I#($v$), which is the current input number for node $v$. The enumeration algorithm essentially prints the input of each node for all possible values of the vector I#, and thus reduces to an enumeration of all possible I# vectors.

```
INPUTSET = INFER-NETWORK(PROFILES,V,parsimonious)
ENUMERATE-NETWORK (V,INPUTSET)
set I#(v) = 1 for all nodes v of V
while I# ≤ ( |INPUTSET(v1)|,…,|INPUTSET(vn)| ) do
  For all nodes v of V do
      Let input be the I#(v) element of INPUTSET(v)
      print v, input
  done
  next(I#)
done
```

Now that we know how to infer networks through sampling or enumeration the dynamics of these network need to be run to 1) verify that the expression profiles given as input can indeed be reproduced by these inferred networks 2) to expand the dynamics beyond the times series that were provided as input, and 3) to predict expression profiles under initial conditions different than those provided. Of particular interest is to compute the steady state or equilibrium dynamics of the networks, these steady states should determine the final response of the networks under the initial conditions they are subjected to. With Boolean networks and Boolean dynamics, these steady states are named attractors (Kauffman 1969). An attractor is a cyclic pattern of expression profiles in which the dynamics is drawn too. Assuming the profiles are given up to a predefined time T, the following routine returns the time step an attractor is found, that is a time $t_1$ where all the nodes have the same expression profiles than at time T.

```
ATTRACTOR (profile,V,T)
For t1 = T-1 to 0 do
    For all nodes v in V do
       if profile(v,t1) ≠ profile(v,T) then end loop fi
    done
    if profile(v,t1) = profile(v,T) then end loop fi
done
if profile(v,t1) = profile(v,T)
   then return(t1)
   else return(∞)
fi
```

The routine given below run the dynamics of a given network for the provided initial conditions and print the expression profiles of all the nodes up to the given time. When the input time is infinite the routine runs the network until an attractor is reached. In the present case the set PROFILES comprise initial expression profiles for all the nodes at time t=0 for various initial conditions. The set *input* comprises for each node *v* its input nodes and associate Boolean function as printed by the routines SAMPLE-NETWORK or ENUMERATE-NETWORK.

```
RUN-NETWORK (PROFILES, V, input, T)
For all profile in PROFILES do
  t1 = 1
  For t = t1 to T do
     For all node v in V do
       let v1,…,vk,f be the input nodes and
       associated Boolean function for node v
       as recorded in input(v)
       profile(v,t) = f(profile(v1,t-1),…,profile(vk,t-1))
     done
```

```
      if (T = ∞) then
         t1 = ATTRACTOR(profile(v1),…,profile(vn),V,t)
         if (t1 < ∞) then t = ∞ fi
      fi
   done
   For all node v in V do
       print profile(v,t1),…,profile(v,t)
   done
done
```

### 5.2.9. Computational complexity of inference algorithms

We provide in this section an analysis of the computational complexity of the algorithms outlined above showing that for gene regulatory networks the algorithms are generally efficient and can be run in polynomial time. Let $n$ be the number of nodes in the networks to be inferred, and let $k$ be the maximum number of input a node can take. While theoretically $k$ is not bounded and can be equal to $n$, because gene regulatory networks follow power law (Basso et al. 2005) with exponent greater than 2 for 100 nodes $k$ is no greater than 5, additionally if parsimonious networks are inferred our experience demonstrates that $k$ does not exceed 3.

Let P be the number of expression profiles and T be the maximum length of the time courses, we assume P and T to be constant independent of $n$. `INFER_FUNCTION` runs in O(PT) = O(1) time steps. `INFER-NETWORK` which calls `INFER-FUNCTION` for every node and every possible k-tuples of input runs in $O(n\mathrm{PT}n+n\mathrm{PT}n^2+n\mathrm{PT}n^3+\ldots+n\mathrm{PT}n^k) = O(n^{k+1})$ steps. It can easily be seen that `COUNT-NETWORK` and `SAMPLE-NETWORK` runs in O($n$) steps, while `ENUMERATE-NETWORK` runs if O($n$|I#|) where |I#| is the number of possible input vectors (i.e. the number of solutions), note that this number can be exponential for $n$, and thus `ENUMERATE-NETWORK` can run for an exponential time and can output an exponential number of solutions. However, if the

number of solutions, |#I| is polynomial for $n$, then ENUMERATE-NETWORK will run in polynomial time. In any case the time taken between two networks output by ENUMERATE-NETWORK reduces to the time to print a network and increment the vector I# both can be carried in O($n$) steps, thus ENUMERATE-NETWORK prints networks in linear time per output.

To compute attractors one need to find a time step for which all the nodes have the same profile than for the last time step provided. The routine ATTRACTOR thus runs at most O($n$T) steps and O($n$) steps for constant T. RUN-NETWORK computes and prints the profile of every node up to time T. When T is finite the time complexity is thus O($n$T) = O($n$). The procedure is repeated for every profile in PROFILES with an overall time complexity of O($n$PT) = O($n$). When T is infinite the profiles are computed and printed till an attractor is found. Let T = $T_\infty$ be the time step an attractor if found. As before computing and printing profiles is performed in O($n$P$T_\infty$) = O($n T_\infty$) steps. The routine ATTRACTOR runs in O(nT) steps, that routine is called $T_\infty$ times, the complexity is thus O($n T_\infty^2$). The overall time complexity is O($n$P$T_\infty^2$) = O($n T_\infty^2$). Note that in the present case $T_\infty$ is not independent of $n$, but is bounded by $2^n$ the number of possible states of the network. Nonetheless, as shown by Kaufman in a seminal paper (Kauffman 1969) gene regulatory networks have relatively few attractors and these are of short length, precisely for a network of $n$ nodes, the number and the length of attractor is O($n$) rather than being proportional to $2^n$. Following Kaufman findings, the computational complexity of RUN-NETWORK for regulatory networks reduces to O($n^3$).

## 5.3. RESULTS

### 5.3.1. Identification of known IL-2-regulated genes in mouse

To determine the filtering parameters for selecting genes regulated by IL-2, we used 18 known genes that differentially expressed in response to IL-2 stimulation. After IL-2 stimulation for 4 hours, the expression level of these IL-2-regulated genes was examined (Table 5.1), the expression level of each gene was ≥ 100 in fluorescent intensity, ≥1.5-

fold comparing to the control (no IL-2 stimulation), with a present call (P) as determined by Affymetrix Microarray Analysis Suite version 5.0 (MAS 5.0).

Table 5.1. Known genes differentially expressed in response to IL-2 stimulation for 4 hours

| Probe ID (No.) | Gene[1] | Description | Mouse CTLL-2 | | Human T cell[2] | |
|---|---|---|---|---|---|---|
| | | | Fold | std | Fold | std |
| | | ***Control*** | | | | |
| | GAPDH | glyceraldehyde-3-phosphate dehydrogenase | 1.09 | 0.11 | | |
| | Actin | beta-actin | 1.04 | 0.16 | | |
| | | ***Up-regulated*** | | | | |
| 1423100_at | Fos | FBJ osteosarcoma oncogene | 8.40 | 9.73 | | |
| 1417409_at (2) | Jun | Jun oncogene | 1.67 | 0.48 | | |
| 1422938_at | Bcl2 | B-cell leukemia/lymphoma 2 | 2.71 | 0.42 | | |
| 1420887_a_at (3) | Bcl2l1 | Bcl2-like 1 (bclx) | 6.09 | 2.05 | 1.40 | 0.20 |
| 1423006_at (3) | Pim1 | proviral integration site 1 (Oncogene PIM1) | 2.12 | 0.64 | 2.90 | 0.40 |
| 1424942_a_at | Myc | myelocytomatosis oncogene | 19.01 | 3.82 | | |
| 1416122_at (7) | Ccnd2 | cyclin D2 | 2.29 | 0.71 | | |
| 1415907_at | Ccnd3 | cyclin D3 | 2.04 | 0.07 | | |
| 1416492_at | Ccne1 | cyclin E1 | 1.66 | 0.37 | | |
| 1420691_at (2) | Il2ra | Interleukin 2 receptor, alpha | 3.89 | 0.91 | 6.60 | 1.50 |
| 1421034_a_at | Il4ra | interleukin 4 receptor, alpha | 2.54 | 0.47 | 2.1 | 0.30 |
| 1448724_at | CIS | cytokine inducible SH2-containing protein (Cish) | 12.49 | 2.27 | | |
| 1416576_at (3) | CIS3 | cytokine inducible SH2-containing protein 3 | 11.09 | 2.79 | | |
| 1450129_a_at | CIS4 | cytokine inducible SH2-containing protein CIS4 | 3.94 | 0.34 | | |
| 1418309_at | TNF-11b | tumor necrosis factor receptor 11b | 10.38 | 3.29 | | |
| 1453851_a_at | GADD45γ | growth arrest and DNA-damage-inducible 45 gamma | 1.62 | 0.48 | 3.30 | 0.70 |
| 1422115_a_at | Flt3l | fms-related tyrosine kinase 3 ligand | 1.90 | 0.40 | 3.00 | 0.50 |
| 1421963_a_at | Cdc25b | cell division cycle 25 homolog B | 2.99 | 0.31 | -2.00 | 0.30 |
| 1420886_a_at (2) | Xbp1 | X-box binding protein 1 | 1.62 | 0.35 | 2.80 | 0.50 |
| 1420669_at | Arnt2 | aryl hydrocarbon receptor nuclear translocator 2 | 8.44 | 6.80 | | |
| 1425947_at | Ifng | interferon gamma | 7.79 | 7.24 | | |
| 1449317_at | Cflar | CASP8 and FADD-like apoptosis regulator | 1.85 | 0.26 | | |
| 1421173_at (2) | IRF4 | interferon regulatory factor 4 | 4.28 | 1.74 | 3.50 | 0.80 |
| 1418518_at | Furin | furin (paired basic amino acid cleaving enzyme) | 4.42 | 0.80 | 2.10 | 0.30 |
| | | ***Down-regulated*** | | | | |
| 1421469_a_at (2) | Stat5a | signal transducer and activator of transcription 5A | -2.03 | 0.24 | | |
| 1422102_a_at (2) | Stat5b | signal transducer and activator of transcription 5B | -1.83 | 0.32 | -3.70 | 1.20 |
| 1419497_at (2) | Cdkn1b | cyclin-dependent kinase inhibitor 1B (P27kip1) | -4.30 | 4.40 | -2.00 | 0.40 |
| 1422734_a_at | Myb | myeloblastosis oncogene | -1.47 | 0.34 | | |
| 1437626_at | Zfp36l2 | zinc finger protein 36, C3H type-like 2 | -2.15 | 0.24 | -3.00 | 0.60 |
| 1449310_at | Ptger2 | prostaglandin E receptor 2 (PGE2-R) | -1.87 | 0.11 | 8.80 | 2.10 |
| 1436861_at | IL-7 | interleukin 7 | -2.81 | 0.48 | | |
| 1450381_a_at | BCL-6 | B-cell leukemia/lymphoma 6 | -10.20 | 9.98 | | |

**Notes:** 1. Blue labeled were genes known differentially expressed in mouse T cell CTLL-2 in response to IL-2 stimulation. Black labeled were genes from literature differentially expressed in human or mouse T cell 2. Data was from Beadling and Smith (2002)

An early study of human T cells indicated that expression of 47 genes changed ≥2 fold in response to IL-2 stimulation for 4 hours (Beadling and Smith 2002), 12 of them have homologous in our differentially expressed genes in response to IL-2 stimulation, with similar expression trend to that in human T cells (Table 5.1). Therefore the filter parameters for selecting genes in response to IL-2 stimulation from the 12 time series points are as follows: (1) present call (P), (2) change call (I, increase, D, decrease), (3) signal intensity ≥100, and (4) ≥1.5-fold change in expression level comparing to the control. 5804 probes were selected according to the filter criteria with ≥1.5 fold change at expression level at least on one time point.

Using tools at Affymetrix NetAffx Analysis Center (http://www.affymetrix.com/analysis/index.affx), the 5084 probes (11.27% of the total 45119 probes on the array) were analyzed according GO categories (Figure 5.1a). 2820 probes (representing 1988 genes) have annotated GO biological processes, 3132 probes (representing 2227 genes) have annotated GO molecular function, and 2719 probes (representing 1941 genes) have annotated GO cellular location. In each category, our selected probes account approximately 13% of the total annotated probes (Figure 5.1a). However, only 1967 probes (representing 1390 genes) are known in three GO categories. Surprisingly, ~ 10% of the probes are with no information for their biological process, molecular function and cellular location, clustering analysis of our gene expression data may provide clues to their function. Details of the GO biological process, molecular function and cellular location of the genes with known GO categories are shown in Figure 5.1b.

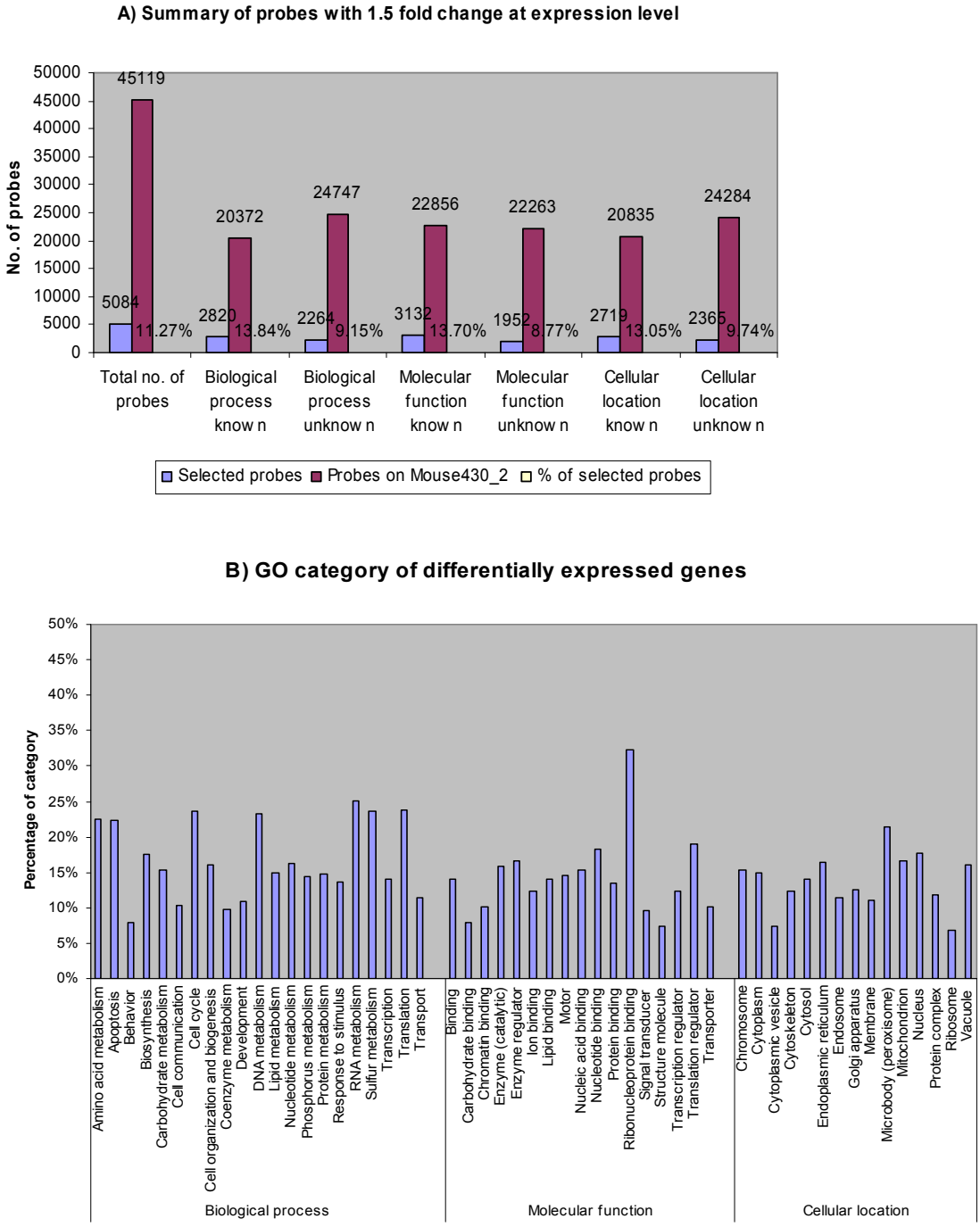


Figure 5.1 GO categories of selected 5804 probes. a) (above), red bars represent total number of probes on Mouse 430_2 array, blue bars represent the number of probes selected and the percentage of the selected probes vs total probes. b) (below), percentage of selected probes vs those on the array for each known GO category in biological process, molecular function and cellular location.

### 5.3.2. Clustering and discretization

The microarray dataset analyzed included 12 time points with 5085 genes at each time step. Two subsets of the microarray dataset were analyzed. The first was the mean-subtracted log of the averaged adjusted profiles, and the second was the mean-subtracted log of the averaged adjusted profiles restricted to transcription factors.

We performed discretizations of these the full dataset and also examined the restriction of the result to transcription factors. As described in Methods, we first computed the appropriate value of $k$ to use in $k$-means. In Figure 5.2, we show the value of the mean internal consistency versus $k$, for $k = 2,\ldots,40$.

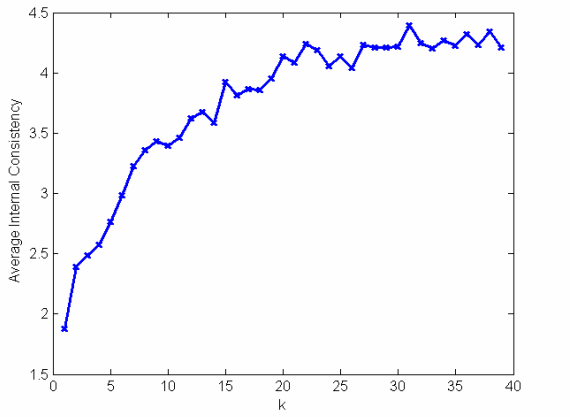


Figure 5.2. Here we show the behavior of the average internal consistency as a function of $k$. Based on the curve we selected $k = 23$ to use with $k$-means.

The internal consistency increases until $k$ is approximately 25 then flattens out. We chose a small local peak at $k = 23$. Using $k$-means with $k = 23$, we obtained the partition of the full dataset shown in Figure 5.3. This clustering also included at least one gene in each cluster when restricted to transcription factors.

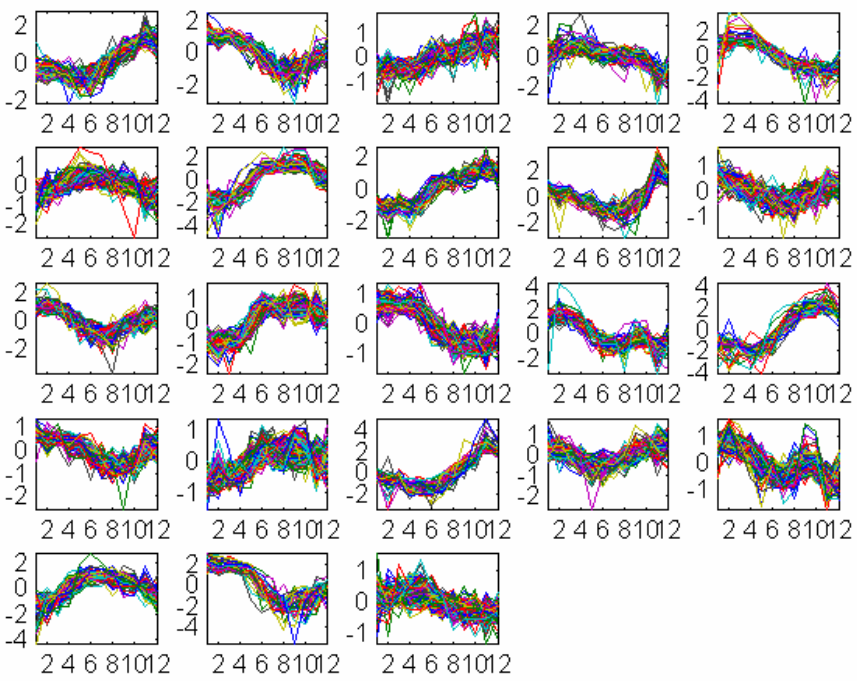

Figure 5.3. This figure shows the 23 meta-genes we used in our analysis.

We then computed SVR representations of the meta-gene groups as described in Methods. Two examples are shown in Figure 5.4.

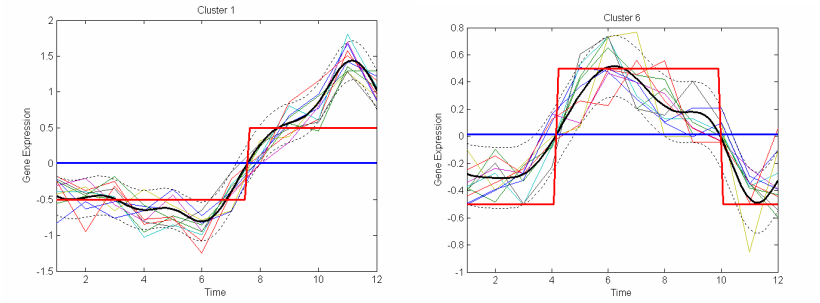


Figure 5.4. Here we show some the results of our discretization method applied to meta-gene groups 1 and 6.

The SVR representations were not computed using every gene in each meta-gene group. To improve computational efficiency, we used only the 10 time courses closest the cluster centers to obtain the SVR curves. The SVR representations were discretized by comparison with the average expression value of the representation. The result of the discretization procedure applied to all 23 meta-genes in given in Table 5.2. The set of 23 discretized profiles reduces to only 12 different profiles as shown in Table 5.3.

Table 5.2. Discretized profiles for the 23 meta-genes of Figure 5.3. Profiles of Figure 5.3 are labeled C1 to C23 reading the figure from left to right and top to bottom. To the cluster number a name is added selecting one arbitrary transcription factor belonging to the cluster. The profiles are reported in a binary notation, 1 is for up-regulated and 0 in for down-regulated.

| | time (h) | | | time (h) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| name | 0 | 0.2 | 0.5 | 1 | 2 | 4 | 6 | 8 | 10 | 12 | 16 | 24 |
| C1-Mybl2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| C2-Ssbp2 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| C3-Csda | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| C4-Jun-Fos | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| C5-Mad4 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| C6-Bcl3 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| C7-Rpo1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| C8-Orc2 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| C9-Foxm1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| C10-Cdkn2c | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| C11-Bcl6 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| C12-Hnr | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| C13-Stat5a | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| C14-Fos | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C15-Mcmd | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| C16-Stat1-6 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| C17-Papola | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| C18-Brca1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| C19-Nsbp1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| C20-Stat5b | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C21-Myc | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| C22-Hdac5 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C23-Stat2 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 5.3. Non identical discretized profiles for the 23 meta-genes of Figure 5.3. The names are those of Table 5.2, to which letter E,L and I have been added. E stands for early genes up-regulated after 1 hour, I for intermediate genes up-regulated after 2 hours, and L for late genes up-regulated after 8 hours.

| | time (h) | | |
|---|---|---|---|
| name | 0 | 0.2 | 0.5 |
| L-Mybl2 | 0 | 0 | 0 |
| L-Mcmd | 0 | 0 | 0 |
| I-Rpo1-Hnr | 0 | 0 | 0 |
| I-Bcl3 | 0 | 0 | 0 |
| I-Myc | 0 | 0 | 0 |
| L-Foxm1 | 1 | 1 | 1 |
| L-Nsbp1 | 1 | 1 | 1 |
| E-Cdkn2c | 1 | 1 | 1 |
| E-Stat5b | 1 | 1 | 1 |
| E-Stat1-6 | 1 | 1 | 1 |
| E-Stat5a | 1 | 1 | 1 |
| E-Jun-Fos | 1 | 1 | 1 |

| time (h) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 4 | 6 | 8 | 10 | 12 | 16 | 24 |
| 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |

Finally, we examined the robustness of our discretization relative to the random starting conditions for *k*-means. We let *k* vary from 2 to 40 and we re-started *k*-means 25 times for each value of *k*. We then computed the average pairwise intersection measure for each of the 25 discretizations to obtain an average intersection value for each *k*. The resulting curve is shown in Figure 5.5.

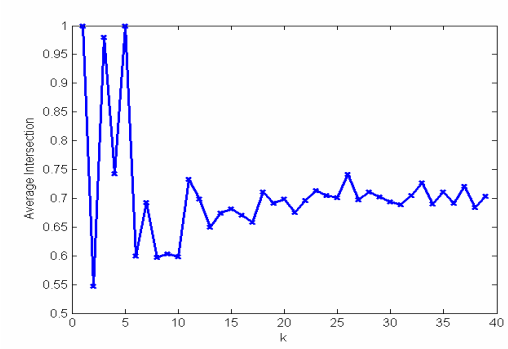


Figure 5.5. This figure shows the average intersection value of 25 random re-starts of *k*-means for each value of *k*. For $k = 23$, the average intersection value is 0.7.

As seen in the above plot, our discretizations are relatively stable and have high intersection for most values of $k$, with some initial fluctuations. This confirms that our choice of $k = 23$ is reasonable for our application.

### 5.3.3. Network Inference

The data listed in Table 5.3 were used as input to the inference algorithms described in the Materials and Methods section. Table 5.3 reveals that the expression profiles of the gene clusters remain constant at 0 hour, 15 min and 30 min. Essentially before 1 hour, the expression profiles of the genes represent the state of the cell under IL-2 starving conditions. At 1 hour and up to 24 hours the expression profiles changes reflect IL-2 addition. We thus have two sets of profiles, IL-2 starved represented by the 3 time points before 1 hour, and IL-2 stimulated given by the 9 time points starting at 1 hour and ending at 24 hours. The routines INFER-NETWORK and ENUMERATE-NETWORK were run using the above two sets of profiles. A total of 161,558 networks were generated. Some of the networks had similarities, for instance it was found that cluster L-Mybl2 is always an activator of cluster L-Nsbp1. Figure 5.6 lists all the activation and inhibition relationships occurring in more than 25% of the networks.

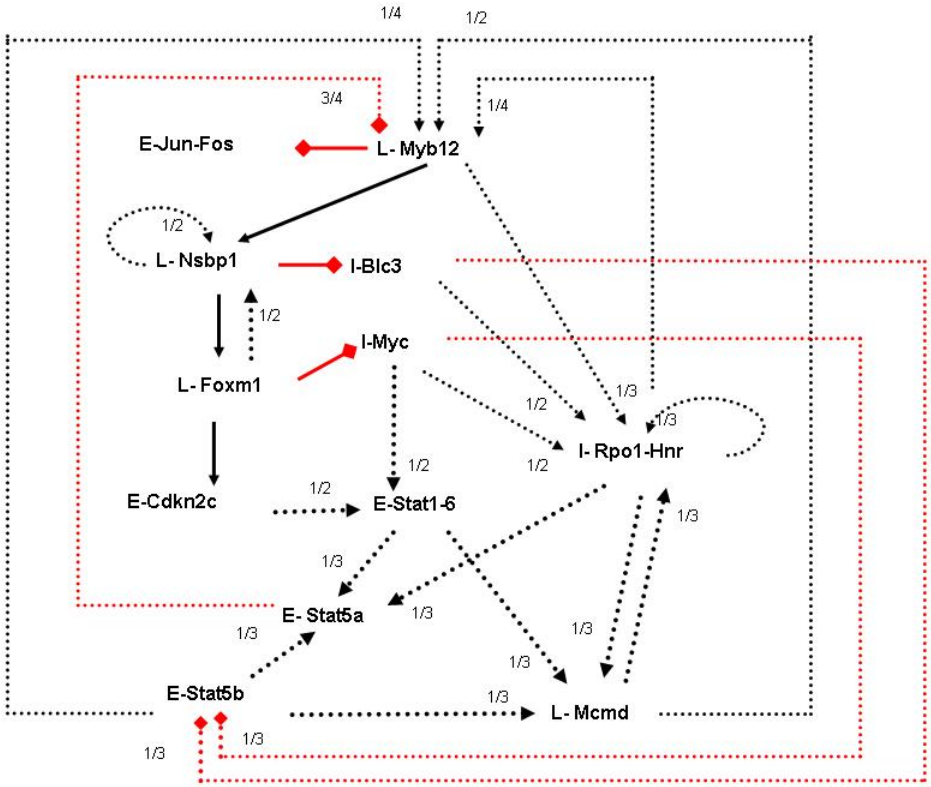


Figure 5.6. This figure gives the activation (black) and inhibition (red) relationships between the 12 clusters of genes of Table 5.3. Solid arrows are for relationships occurring in all networks. The numbers associated with the dashed arrows indicate the fraction of networks (among 161,568) having that relationship. Only relationships occurring in more than 25% of the network are depicted.

The dynamics of all the 161,568 networks were analyzed and in particular attractors were determined for the two initial conditions, IL-2 starved, and IL-2 stimulated. We recall from the Materials and Methods section that an attractor is a cyclic pattern where the dynamics is drawn to. It was found that 160,657 (99.4%) of these networks had a single point steady state dynamics for the IL-2 stimulated condition as shown in Table 5.4.

Table 5.4. Steady state dynamics for 99.4% of T cell inferred networks (160,567 out of 161,568). For both conditions IL-2 starved and IL-2 stimulated the steady state dynamics is a fixed point, i.e., the expression profiles remain fixed at t0 for IL-2 started, and fixed at t4 for IL-2 stimulated.

| | time (h) | |
|---|---|---|
| name | 0 | t0 |
| L-Mybl2 | 0 | 0 |
| L-Mcmd | 0 | 0 |
| I-Rpo1-Hnr | 0 | 0 |
| I-Bcl3 | 0 | 0 |
| I-Myc | 0 | 0 |
| L-Foxm1 | 1 | 1 |
| L-Nsbp1 | 1 | 1 |
| E-Cdkn2c | 1 | 1 |
| E-Stat5b | 1 | 1 |
| E-Stat1-6 | 1 | 1 |
| E-Stat5a | 1 | 1 |
| E-Jun-Fos | 1 | 1 |

| time (h) | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 24 | t1 | t2 | t3 | t4 | t4 |
| 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |

After IL-2 addition, networks leading to fixed point dynamics should not be considered as the T cell follow a cell cycle leading to proliferation and differentiation. A fixed point dynamics is a state where all the genes remains at the same expression level, yet live T cells should undergo cell cycle comprising Mitosis, Interphase, Regulation and Cell Division, and gene expression should fluctuate. Discarding all networks having a fixed point steady state dynamics when IL-2 is added, only 0.6% (911 out of 161,568) networks remain. As depicted in Table 5.5, these networks undergo a steady state going through 3 times points (t3, t4, and t5).

Table 5.5. Steady state dynamics for 0.6% of T cell inferred networks. For the IL-2 starved the steady state dynamics is a fixed point (t0). For the IL-2 stimulated condition, the steady state dynamic follows a cyclic pattern going through 3 steps, t3, t4, and t5.

| | time (h) | |
|---|---|---|
| name | 0 | t0 |
| L-Mybl2 | 0 | 0 |
| L-Mcmd | 0 | 0 |
| I-Rpo1-Hnr | 0 | 0 |
| I-Bcl3 | 0 | 0 |
| I-Myc | 0 | 0 |
| L-Foxm1 | 1 | 1 |
| L-Nsbp1 | 1 | 1 |
| E-Cdkn2c | 1 | 1 |
| E-Stat5b | 1 | 1 |
| E-Stat1-6 | 1 | 1 |
| E-Stat5a | 1 | 1 |
| E-Jun-Fos | 1 | 1 |

| time (h) | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 24 | t1 | t2 | t3 | t4 | t5 | t3 |
| 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |

The set of networks corresponding to the steady state dynamics of Table 5.5 is given in Figure 5.7.

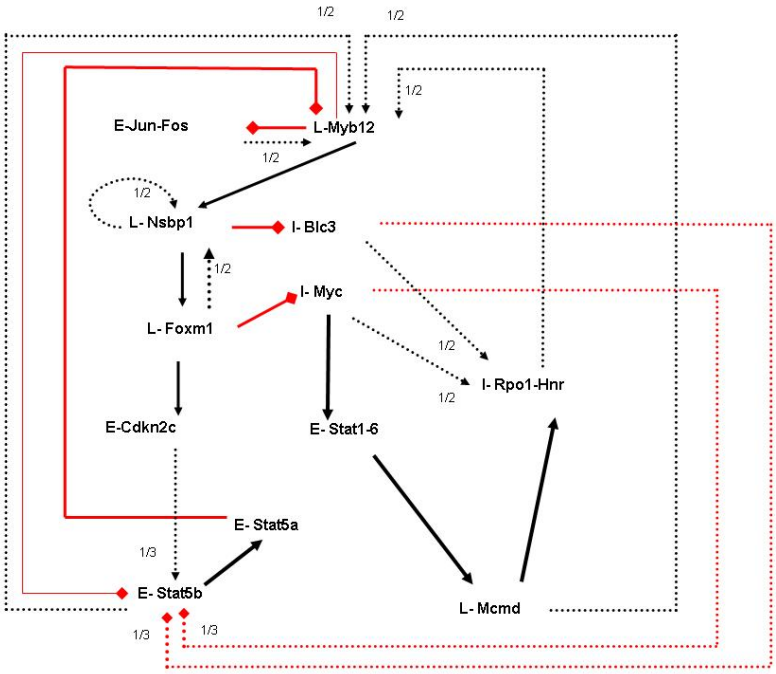


Figure 5.7. This figure gives the activation (black) and inhibition (red) relationships between the 12 clusters of genes following the dynamics of Table 5.5. Solid arrows are for relationships occurring in all networks. The numbers associated with the dashed arrows indicate the fraction of networks (among 911) having that relationship.

## 5.4. DISCUSSION

According to the expression profiles of the 5804 genes in the 12 time points of Tables JLF-1 and JLF-2, the dynamical fingerprint defining the immune response of IL-2 stimulation was determined and is composed of four classes. 1) Early response gene (up-regulated after 1 hour) composed of clusters E-stat5b, E-stat5a, and E-Jun-Fos. 2) Early cyclic response genes (up-regulated after 1 hour and again 16 hours) comprising clusters E-Cdkn2c and E-Stat1-6. 3) Intermediate response genes (up-regulated after 2 hours) composed of clusters I-Rpo1-Hnr, I-Bcl3, and I-Myc. 4) Late response genes (up-regulated after 8 hours) comprising clusters L-Mybl1, L-Mcmd, L-Foxm1, and L-Nsbp1. The viable networks inferred from the 12 time points series and depicted in Figure 5.7 reveals that in general early genes activate other early genes and late genes. Intermediate genes activates late genes and inhibit early genes. Note that early genes inhibited by intermediate genes can be up-regulated when the intermediate genes are down-regulated. Late genes activate other late genes and inhibit early and intermediate genes, note again that early and intermediate genes inhibited by late genes can be up-regulated when late gene are down-regulated. While these relationships are consistent with Boolean logic, to further elucidate the relationships between genes and cluster of genes we report next a throughout annotation analysis of the four classes mentioned above.

The number of probes varies in each of the four classes, 2523 probes in class 1 representing 1904 early response genes (including cyclic response genes), 1935 probes in class 3 representing 1400 intermediate response genes, 438 probes in class 4 representing 365 late response genes. In each class, more than half of the probes were characterized by their GO biological process, molecular function and cellular location (Figure 5.8). We next discuss genes in each class function in such biological processes as immune response, cell cycle, signal transduction and transcription.

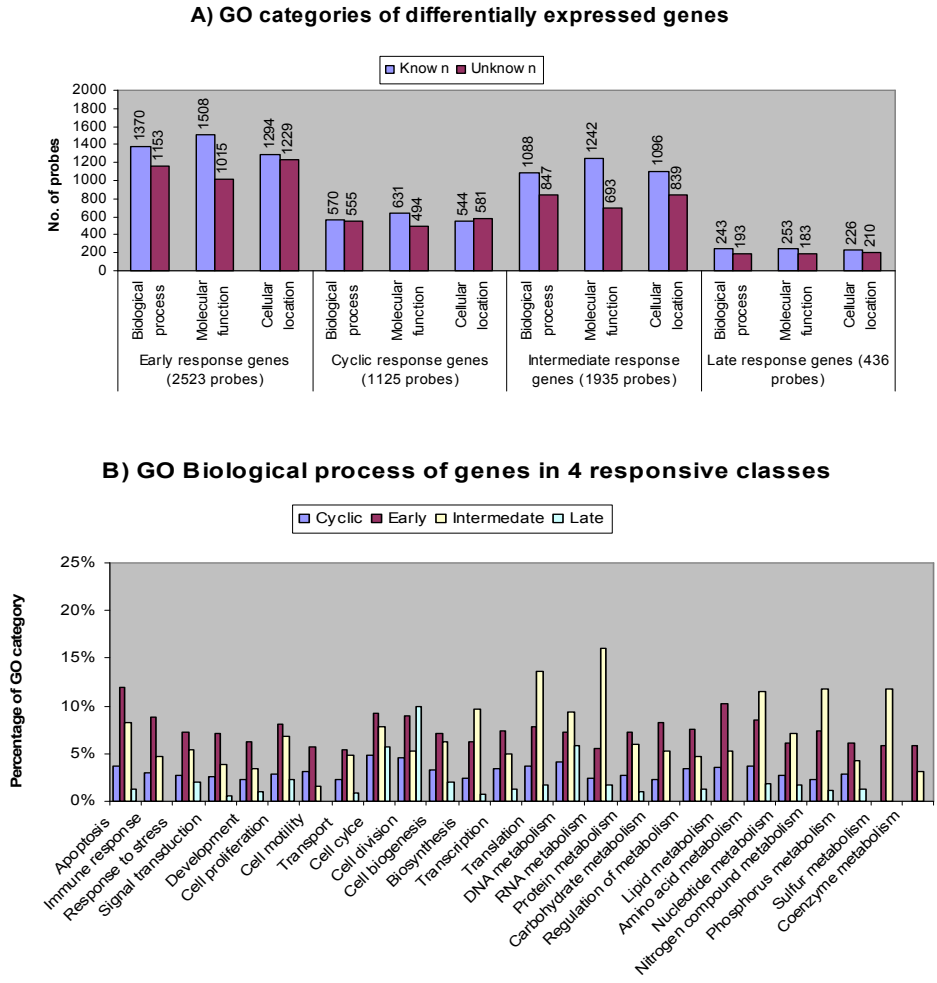


Figure 5.8. GO categories of genes in four classes. A (above), in each class, blue bar represent the number of selected probes with characterized GO category, red bars represent probes with unknown GO category. B (below), percentage of selected probes vs those on the array of the known GO category in biological process.

### 5.4.1. **Early response genes** (2523 probes)

Genes in this category responded to IL-2 after 1 hour stimulation, promoting T cell proliferation by up or down regulating relevant genes involved in various biological processes. There are 2523 probes in this category from 11 clusters (out of 23 clusters), 1125 probes of them from 4 clusters are also defined as cyclic response genes since they were up-regulated at 1 hour and then again at 16 hour (see below). Here we will present

the results of rest 1140 probes in this category, the biological processes of 805 probes representing 643 genes were well characterized.

**5.4.1.1. Immune response**. There are 42 early response genes represented by 53 probes involve immune response (Figure 5.9). Specifically, Stat5a, Stat5b, Bcl10, CD28 antigen, Tcrb-V13 (T-cell receptor beta, variable 13), Ncf1 (neutrophil cytosolic factor 1) and Dock2 (dedicator of cyto-kinesis 2) function in cellular defense response. Stat5a, Stat5b, Ncf1, and Dock2 also function in inflammatory response, as well as Tnfrsf1b (tumor necrosis factor receptor superfamily, member 1b), Rac1 (RAS-related C3 botulinum substrate 1), Nfkbiz (nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, zeta), Fcgr3 (Fc receptor, IgG, low affinity III) and complement component 3. We also find 15 genes involve in immune cell activation, these genes are Stat5a, Stat5b, Dock2, Vav1 oncogene, CD28 antigen, Lcp2 (lymphocyte cytosolic protein 2), Inpp5d (inositol polyphosphate-5-phosphatase D), Igbp1 (immunoglobulin (CD79A) binding protein 1), Ap3b1 (adaptor-related protein complex 3, beta 1 subunit), Prkcd (protein kinase C, delta), Mink1 (misshapen-like kinase 1), Egr1 (early growth response 1), Rgs1 (regulator of G-protein signaling 1), Cd3e (CD3 antigen, epsilon polypeptide) and Fcgr3 (Fc receptor, IgG, low affinity III). Additionally, Stat5a, Stat5b, Fcgr3, Cd28, Rnf128 (ring finger protein 128) and Inpp5d (inositol polyphosphate-5-phosphatase D) have function in cytokine synthesis, and 4 genes including H2-K1 (histocompatibility 2/K1, K region) and Fcgr3 have function in antigen processing. Some of the genes described also involve in regulation of immune response.

Interestingly, among the genes involve in various immune responses, only 14 genes are involve only one immune response, other genes such as Stat5a, Stat5b and Fcgr3 involve in 5 immune responses, respectively; CD28, Inpp5d and Dock2 involve in 4 immune responses, respectively, Ap3b1 and complement component 3 involve 3 responses, Bcl10, Cd3e, Rnf128, Prkcd and Ncf1 involve 2 response, suggesting they have very critical function in mouse T cell immune responses. Finally, 15 immune response genes have no function in the described responses , such as Stat3, Fasl (Fas ligand), Irf2 (interferon regulatory factor 2), lymphotoxin B, lymphocyte-activation gene 3, Ccl27

(chemokine (C-C motif) ligand 27), oncostatin M, and Mr1 (major histocompatibility complex, class I-related).

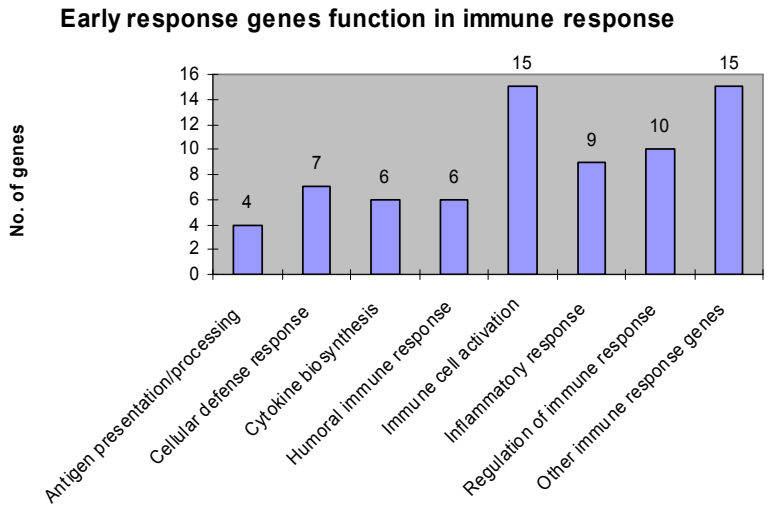


Figure 5.9. Early response genes function in immune response. The number on top of each bar is the number of genes involved in that response.

**5.4.1.2. Cell cycle**. In contrast, more genes (55 genes represented by 71 probes) function in cell cycle (Figure 5.9). Cyclin T2, Cdkn1b (cyclin-dependent kinase inhibitor 1B (P27)), Camk2g (calcium/calmodulin-dependent protein kinase II gamma) and Ppp3ca (protein phosphatase 3, catalytic subunit, alpha isoform) control G1/S transition in interphase of the cell cycle, this transition ultimately determines if a cell is going to divide or not, thus is critical for cell proliferation. More genes (8 genes) involve in mitosis, these genes encode cyclin G2, sirtuin 2, cell division cycle 2-like 5 (Cdc2l5), MAD2L1 binding protein (Mad2l1bp), pituitary tumor-transforming 1 (Pttg1) and katanin p60 (ATPase-containing) subunit A1 (Katna1), neural precursor cell expressed, developmentally down-regulated gene 9 (Nedd9), microtubule-associated protein RP/EB family/member 2 (Mapre2). Even more genes (19 genes) involve in regulation of cell cycle, Jun, Fos, Stat5a, Stat5b, Cdkn1b (cyclin-dependent kinase inhibitor 1B (P27)) are well known genes which also have function in signal transduction and/or transcription. Others like Bcl10, cyclin G2, cyclin I, cyclin T2, ELK3 (member of ETS oncogene family), Map3k8 (mitogen activated protein kinase kinase kinase 8) and Cdc2l5 (cell

division cycle 2-like 5 (cholinesterase-related cell division controller)) also function in other biological processes.

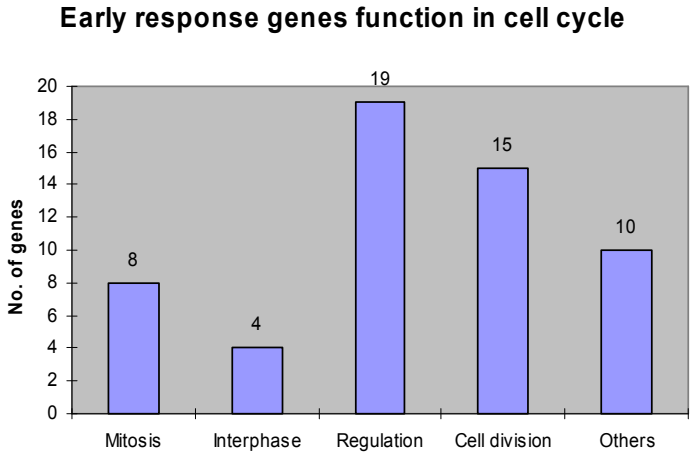


Figure 5.10. Early response genes involved in cell cycle. The number on top of each bar is the number of genes involved in that response.

**5.4.1.3. Signal transduction**. We find 129 genes (represented by 169 probes) to function in signal transduction, involving 12 signaling transduction pathways (Figure 5.11). Among the up-regulated are genes regulating signal transduction by a variety of cytokines. Janus kinase *Jak1* and *Jak3* initiate IL-2 signal transduction pathways by docking protein Jak1 and Jak3 on IL-2 receptor γ and β chain respectively. The CIS (cytokine-inducible SH2-containing protein) family gene *CIS2, CIS3*, also known as the *SOCS2* and *SOCS3* (suppressors of cytokine signaling) gene, are immediate-early gene responding to IL-2 and function in regulating IL-2 signal transduction. Well known nuclear transcription factor gene *Fos* and *Jun* were also up-regulated, the JUN protein directly activate gene transcription in response to cell stimulation, the FOS protein cooperates with the JUN product in fostering gene transcription (Marx 1988).

Specifically 4 genes function in MAPKKK cascade, as expected, mitogen activated protein kinase kinase kinase 7 (Map3k7) and mitogen-activated protein kinase 8 interacting protein 3 (Mapk8ip3) are in this group. 5 genes involve in JAK-STAT cascade, such as Janus kinase 2, signal transducer and activator of transcription of Stat3, Stat5a, Stat5b and oncostatin M. Stat5a and Stat5b also have function in various immune

responses and other signal transduction pathways, suggesting that they have a critical role in mouse T cells in response to IL-2 stimulation. There are 21 genes function in small GTPase mediated signaling transduction, such as Rho GTPase activating protein 1 (Arhgap1) and IQ motif containing GTPase activating protein 1 (Iqgap1) gene. Interestingly, 12 RAS related genes, Ras homolog gene family member C, B, G and Q, RAS oncogene family 20, 27b and 31, RAS-related genes Rac1, Rac2 and Rrad, v-ral simian leukemia viral oncogene homolog (Rala) and Harvey rat sarcoma oncogene subgroup R (Rras), involve in the pathway, suggesting that RAS related genes are significantly important in the pathway. Additionally, 16 genes function in G-protein coupled signaling pathway, these genes encode 3 regulators of G-protein signaling (regulator 1, 2, 11), 4 G-protein guanine nucleotide binding proteins (alpha 15, beta 5, gamma 2 and gamma11), 4 receptors such as gastric inhibitory polypeptide receptor, gamma-aminobutyric acid (GABA-B) receptor 1, cysteinyl leukotriene receptor 2 and coagulation factor II (thrombin) receptor-like 3, as well as calmodulin 3 and adenylate cyclase 7.

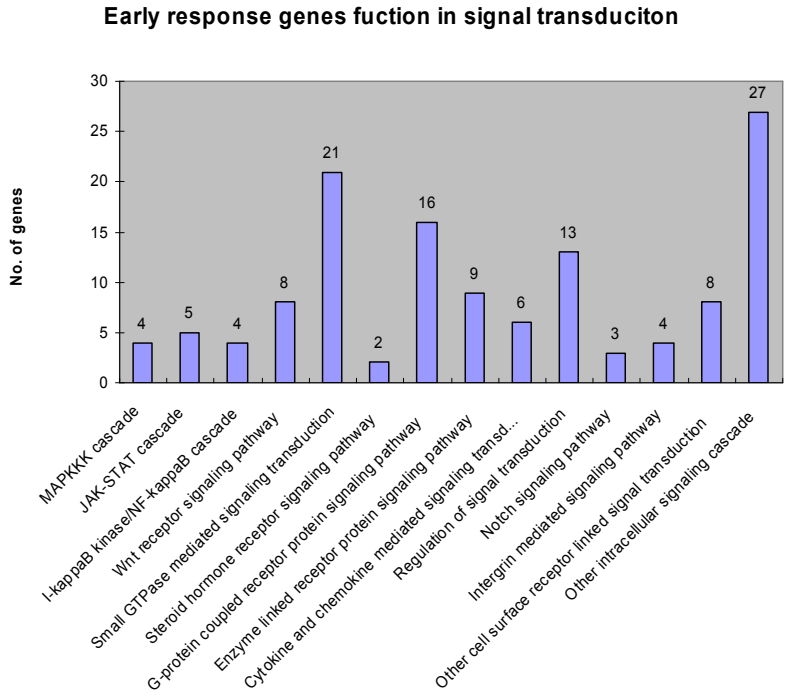


Figure 5.11. Early response genes involved in signal transduction pathways. The number on top of each bar is the number of genes involved in that response.

Furthermore, 8 genes function in other cell surface receptor linked signal transduction, they are interleukin receptor gene Il1r2 (interleukin 1 receptor, type II), Il10ra (interleukin 10 receptor, alpha) and Il15ra (interleukin 15 receptor, alpha). Additionally 27 genes are involved in other intracellular signaling cascades, such as well known signaling protein gene Jak1 (Janus kinase 1), Jak3 (Janus kinase 3), Itk (IL2-inducible T-cell kinase), Fyn (Fyn proto-oncogene), Lcp2 (lymphocyte cytosolic protein 2) and Stat2 as well as protein kinase C (delta, eta and nu). Genes from this group also function in I-kappaB kinase/NF-kappaB cascade, wnt receptor signaling pathway, steroid hormone receptor signaling pathway, enzyme linked receptor protein signaling pathway, cytokine and chemokine mediated signaling transduction, notch signaling pathway and intergrin mediated signaling pathway as well as regulation of signal transduction (Figure 5.11).

**5.4.1.3. Transcription.** There are 109 genes (represented by 145 probes) function in transcription. 77 of them encode transcription factor, some are familiar such as Jun, Fos, Stat2, Stat3, Stat5a, Stat5b, Bcl6, zinc finger protein gene Zfp99, zfp198, zfp336, early growth response 1, and 2 (Egr1, Egr2), cyclin T2, serum response factor (Srf), fos-like antigen 2 (Fosl2), interferon regulatory factor 2 (Irf2), Rnf14 (ring finger protein 14), but majority of them were newly identified transcription factors regulated by IL-2.

**5.4.2. Cyclic response genes** (1125 probes)

In this category, there are 1125 probes with >1.5 fold change at expression level, these genes also belong to early response gene class but not included in the description above. The GO biological processes of 570 probes (representing 471 genes) were well characterized, these genes involve in various biological processes (Figure 5.8).

**5.4.2.1. Immune response.** We find 22 genes having function in immune response (Figure 5.12). 2 genes involve in cellular response, they are Irf1 (interferon regulatory factor 1) and Lyst (lysosomal trafficking regulator). More genes involve in regulation of immune response as well as in immune cell activation, they are interleukin 7, T-box 21/zeta-chain (TCR), high mobility group box 3, tumor necrosis factor superfamily member 13b (Tnfsf13b), and those encoding associated protein kinase Zap70, interferon

regulatory factor 1 (Irf1), inositol 1,4,5-trisphosphate 3-kinase B (Itpkb), adaptor-related protein complex 3 beta 1 subunit (Ap3b1) and delta 1 subunit (Ap3d1).

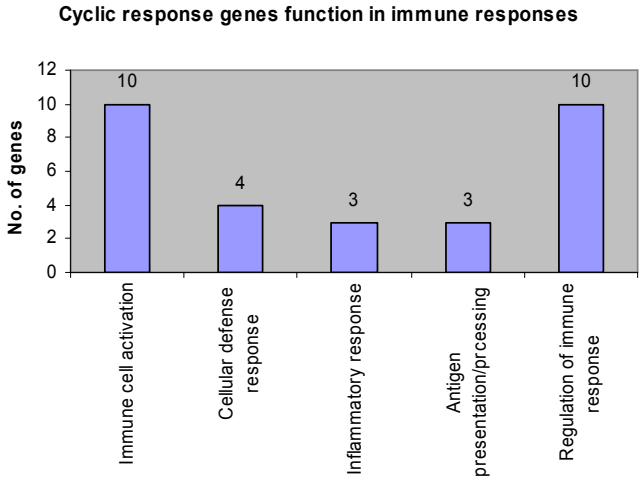


Figure 5.12. Cyclic response genes function in immune response. The number on top of each bar is the number of genes involved in that response.

**5.4.2.2. Cell cycle**. In the category of cyclic genes, 48 genes represented by 58 probes function in cell cycle (Figure 5.13), adding the 55 cell cycle genes described in early response genes' class, thus there are more than 100 early response genes in total function in cell cycle. Specifically, 5 genes function in interphase, checkpoint supressor 1 (Ches1) functions in G2 phase, the delta isoform of protein phosphatase 1D magnesium-dependent (Ppm1d) functions in G2/M transition, while phosphatase 3, catalytic subunit, alpha isoform (Ppp3ca), calcium/calmodulin-dependent protein kinase II gamma (Camk2g) and nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 1 (Nfatc1) function in G1/S transition. More (16) genes function in mitosis, they are genes encode anaphase promoting complex subunit 7 (Anapc7), cell division cycle 25 homolog B (Cdc25b), cyclin B2 (Ccnb2), fizzy/cell division cycle 20 related 1 (Fzr1), microtubule-associated protein (RP/EB family, member 2, Mapre2), NIMA (never in mitosis gene a)-related expressed kinase 1 (Nek1), pituitary tumor-transforming 1 (Pttg1), platelet-activating factor acetylhydrolase, isoform 1b (beta1 subunit, Pafah1b1) and stromal antigen 2 (Stag2). There are 21 genes function in regulation of cell cycle, sestrin 3 (Sesn3), Sestrin 1 (Sesn1), cyclin-dependent kinase inhibitor 3 (Cdkn3) and

microtubule-actin crosslinking factor 1 (Macf1) function in cell cycle arrest; Atr and Ches1 are cell cycle checkpoint genes encoding DNA damage checkpoint protein ataxia telangiectasia and rad3 related and checkpoint supressor 1, respectively; while the rest genes regulate cell cycle, such as cyclin-dependent kinase inhibitor 3 (Cdkn3) and B-cell leukemia/lymphoma 10 (Bcl10) and transcription factor Tfdp2, and E4f1. Additionally, septin 4, setpin 6, septin 11, FCH domain only 2 (Fcho2), CDC28 protein kinase regulatory subunit 2 (Cks2) and SLIT-ROBO Rho GTPase activating protein 2 (Srgap2) have function related to cell division. The up-regulation of a lot more genes involve in mitosis, cell division, and regulation of cell cycle suggests that that activation of these genes be required for cell proliferation after IL-2 stimulation.

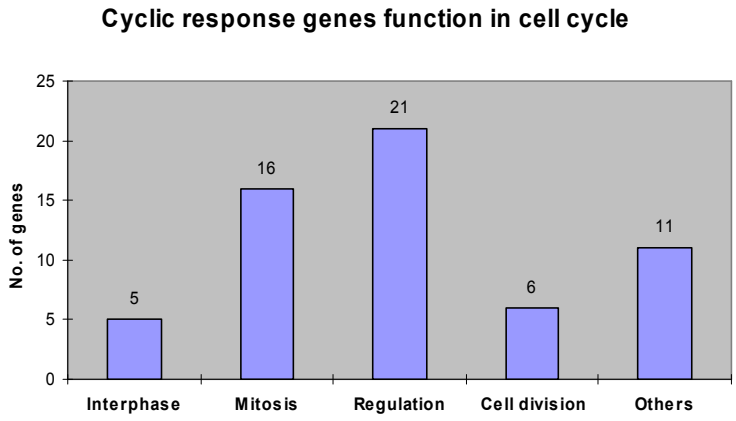


Figure 5.13. Cyclic response genes involved in cell cycle. The number on top of each bar is the number of genes involved in that response.

**5.4.2.3. Signal transduction**. In addition to the 129 early response genes function in signal transduction described, 81 cyclic response genes (represented by 99 probes) are involved in signal transduction (Figure 5.14), therefore there are more than 200 early response genes in total function in signal transduction. Specifically these genes involve in MAPKKK cascade (5 genes), JAK-STAT cascade (3 genes), intracellular signaling cascade (22 genes), I-kappaB kinase/NF-kappaB cascade (5 genes), small GTPase mediated signal transduction (15 genes), cAMP-mediated signaling (4 genes), Wnt receptor signaling pathway (6 genes), G-protein coupled receptor protein signaling

pathway (12 genes), Regulation of signal transduction (6 genes), enzyme linked receptor protein signaling pathway (4 genes) and other signal transduction pathways (15 genes). There are 9 kinase genes: Janus kinase 2 (Jak2), PX domain containing serine/threonine kinase (Pxk), protein kinase C, nu (Prkcn), phosphatidylinositol 3-kinase, regulatory subunit, polypeptide 2 (p85 beta), mitogen activated protein kinase kinase 5 (Map2k5), inositol 1,4,5-trisphosphate 3-kinase B (Itpkb), inhibitor of kappaB kinase beta (Ikbkb), G protein-coupled receptor kinase 5 (Gprk5) and zeta-chain (TCR) associated protein kinase (Zap70). In addition, 7 genes are G-protein related: protein-coupled receptor Edg5, Edg6, Gpr56 and Gpr160, G-protein signaling modulator 3 (Gpsm3), G protein pathway suppressor 2 (Gps2) and regulator of G-protein signaling 3 (Rgs3); and there are 6 RAS oncogene family members: Rab9, Rab8b, Rab8a, Rab4a, Rab37 and Rab19, member of RAS oncogene family-like 4 (Rabl4).

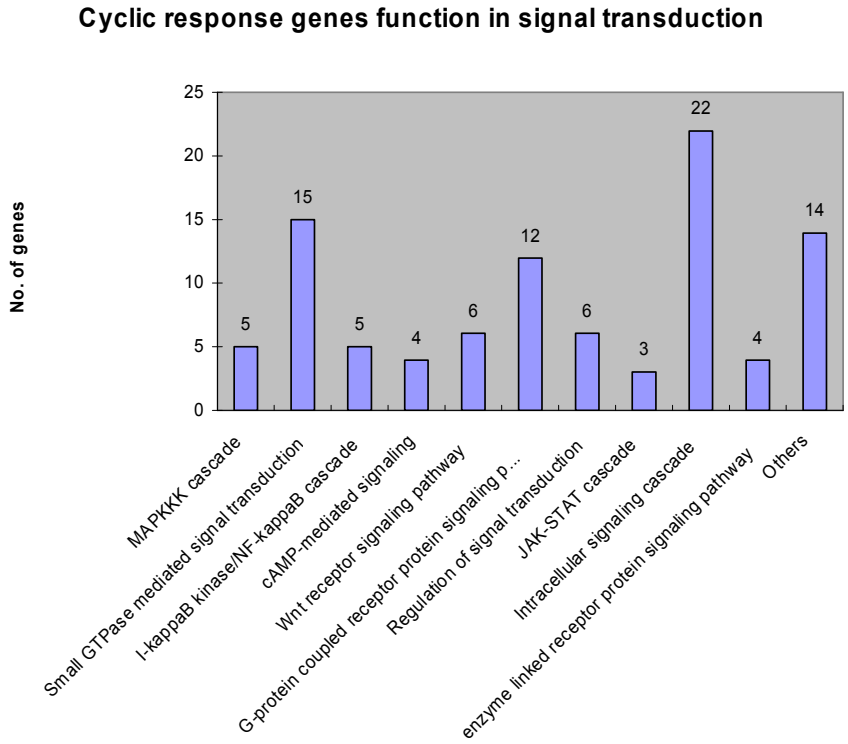


Figure 5.14. Cyclic response genes involved in signal transduction pathways. The number on top of each bar is the number of genes involved in that response.

Additionally, Stat1, Stat6, Rapgef1 (guanine nucleotide exchange factor), Rapgef2 and Arhgef18, Interleukin 17 receptor (Il17r), interleukin 12 receptor, beta 2 (Il12rb2),

chemokine (C-X-C motif) receptor 3 (Cxcr3) and chemokine (C-C motif) ligand 25 (Ccl25), caspase recruitment domain 4 and 6 (card4, card6) are also in this group function in various signal transductions.

**5.4.2.4. Transcription.** There are 95 genes (represented by 126 probes) function in transcription. Therefore, a total of 200 early response genes involve in gene transcription (see also the 105 early response genes described above). Specifically 74 genes (98 probes) encode transcription factors, such as nuclear receptor related Nr3c1, Nr2c2, Ncor2 and Ncoa1, nuclear factor Nfe2l1, B-cell leukemia/lymphoma 6 (Bcl6), cyclin-dependent kinase inhibitor 2C (Cdkn2c), E4F transcription factor 1 (E4f1), interferon regulatory factor 1 (Irf1), nuclear factor of activated T-cells (Nfatc3, Nfatc1, Nfat5), zinc finger proteins (Zfpm1, Zfp99, Zfp592, Zfp277, Zfp207 and Zipro1) and signal transducer and activator of transcription (Stat1, Stat5a and Stat6). The remaining 22 genes (28 probes) encode DNA binding protein such as PHD finger protein Phf14 and Phf21a, leucine-rich repeat kinase 1 (Lrrk1) and zinc finger containing proteins Zfp496, Zbtb8 and Zbtb24. Thus there are more than 150 transcription factor involve in early response to IL-2 stimulation.

**5.4.3. Intermediate response genes** (1935 probes)

Exactly 1400 genes represented by 1935 probes are intermediate response genes, 768 of the genes were characterized in terms of GO biological processes (Figure 5.8). A few genes known responsive to IL-2 stimulation enhance T cell proliferation through various pathways, such as anti-apoptosis gene Bcl2 and Bcl2-like 1, immune response gene Il2ra (interleukin 2 receptor, alpha), Il4ra (interleukin 4 receptor, alpha), Tnf (tumor necrosis factor), Ifng (interferon gamma). Additionally, myelocytomatosis oncogene (Myc), cyclin D2, cyclin D3, cyclin E1 and cyclin E2 as well as 3 cyclin-dependent kinase gene Cdk6, Cdk7 and Cdk8, promote T cell proliferation through regulation of cell cycle, cell division or DNA replication initiation.

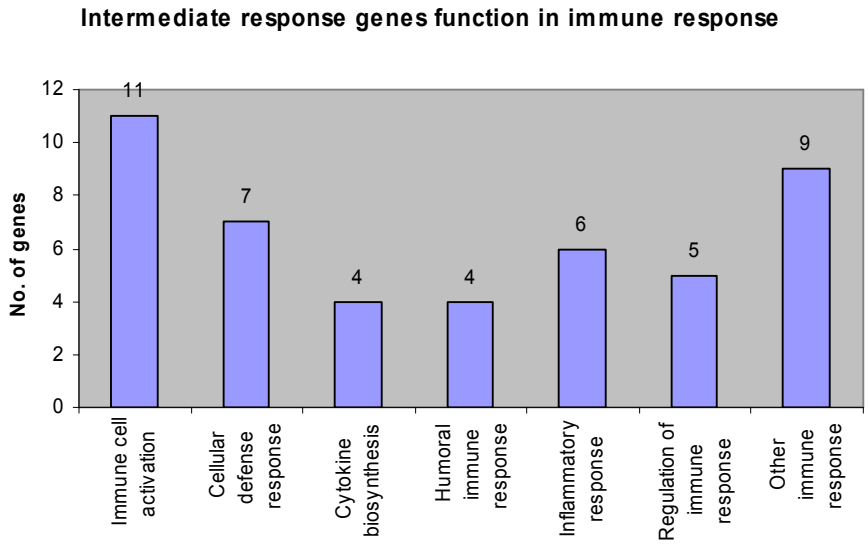


Figure 5.15. Intermediate response genes function in cell cycle. The number on top of each bar is the number of genes involved in that function.

**5.4.3.1. Immune response.** There are 28 genes represented by 38 probes involve in immune response (Figure 5.15). Specifically, Tnf (tumor necrosis factor), Lta (lymphotoxin A), Igj (immunoglobulin joining chain) and C1qbp (complement component 1, q subcomponent binding protein) involve in humoral immune response. Tnf and Lta also function in inflammatory response and cellular defense response. The well known genes Ifng (interferon gamma), Il2ra (interleukin 2 receptor, alpha chain) and Tlr4 (toll-like receptor 4) were also selected as inflammatory response genes. Additional defense response genes are Prkcq (protein kinase C, theta), Nfkb1 (nuclear factor of kappa light chain gene enhancer in B-cells 1, p105), Gadd45g (growth arrest and DNA-damage-inducible 45 gamma) and Cd8a (CD8 antigen, alpha chain). Ifng, Gadd45g, Prkcq and Nfkb1 also involve in cytokine biosynthesis. A total of 11 genes function in immune cell activation that ultimately results in T cell proliferation, 6 of them (Il2ra, Ifng, Prkcq, Gadd45g, Cd8a and Pdcd1lg2 (programmed cell death 1 ligand 2) also function in other immune responses or regulation of immune response, cyclin D3, Flt3l (FMS-like tyrosine kinase 3 ligand), Ndrg1 (N-myc downstream regulated gene 1), Hells (helicase, lymphoid specific) and Bsf3 (Cardiotrophin-like cytokine factor 1) only function in immune cell activation.

Among the additional 9 immune response genes are interleukin related genes Il24 (interleukin 24), Il10 (interleukin 10) and Il4ra (interleukin 4 receptor/alpha), interferon related gene Ifi35 (interferon-induced protein 35) and Icsbp1 (interferon consensus sequence binding protein 1), Gp49a (glycoprotein 49 A), Daf1 (decay accelerating factor 1), Serpina3g (serine (or cysteine) proteinase inhibitor, clade A, member 3G) and Ppp3cb (protein phosphatase 3, catalytic subunit, beta isoform).

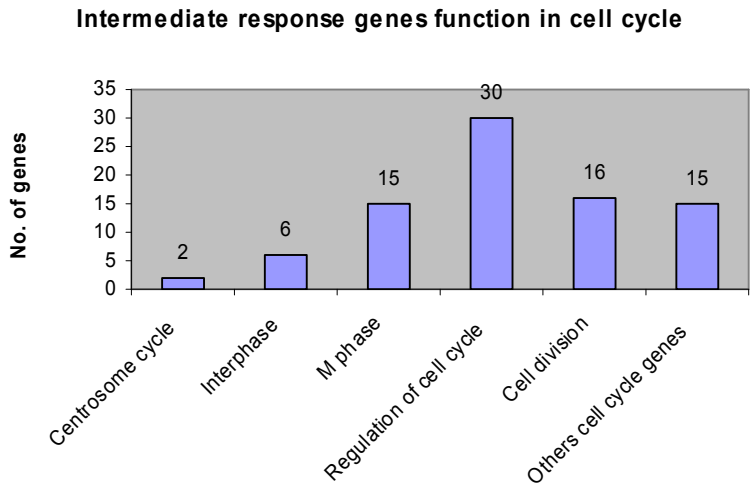


Figure 5.16. Intermediate response genes function in cell cycle. The number on top of each bar is the number of genes involved in that function.

**5.4.3.2. Cell cycle.** The number of intermediate response genes (59 genes represented by 94 probes) involve in cell cycle is approximately half the number of early response genes function in the cell cycle (Figure 5.16). Interphase gene Gspt1 (G1 to S phase transition 1) and Skp2 (S-phase kinase-associated protein 2 (p45)) control G1 to S phase transition; Cdk6 (cyclin-dependent kinase 6), transcription factor Dp1 and DnaJ (Hsp40) homolog subfamily C, member 2 (Dnajc2) function in G1, S and G2 phase, respectively; DNA damage checkpoint kinase 1 (Chek1) is critical in G2/M transition of mitotic cell cycle. In addition, majority of the 15 mitotic (M) phase genes involve in cell division, such as Cdc6 (cell division cycle 6), Cdc16, Cdc25a, Cdca7 (cell division cycle associated 7), chromosome condensation 1 and large tumor suppressor 2. Interestingly, there are more genes (30 genes) with function in regulation of cell cycle, some of them are well known

genes, such as cyclin D2, cyclin D3, cyclin E1 and cyclin E2, as well as Cdk6, Myc, Nmyc1 (neuroblastoma myc-related oncogene 1) and E2f3 (E2F transcription factor 3), others like Tfdp1, Skp2, Chek1, Gspt1, Mphosph6 (M phase phosphoprotein 6), Tnfsf5ip1 (tumor necrosis factor superfamily, member 5-induced protein 1) and Ran (RAS oncogene family member) also function in M or interphase. On the other hand, some cell cycle genes also function in DNA replication, such as Mcm2 (minichromosome maintenance deficient 2), Mcm3, Mcm6 and Mcm7, Pols (DNA polymerase sigma), Cdc6 (cell division cycle 6 homolog), Rpa1 (replication protein A1); some genes involve in DAN repairment, such as RAD51, Hus1, Mlh1 (mutL homolog 1), Chek1 (checkpoint kinase 1 homolog), H2afx (H2A histone family, member X); Cdk7 (cyclin-dependent kinase 7), Cdc7 (cell division cycle 7) and Ptp4a1 (protein tyrosine phosphatase 4a1) involve in protein phosphorylation; Ranbp1 (RAN binding protein 1) functions in spindle organization and biogenesis; Tube1 (epsilon-tubulin 1) is related to cellular morphogenesis.

**5.4.3.3. Signal transduction.** Similarly, there are 96 intermediate response genes represented by 144 probes function in signal transduction, less than half of the early response genes involve in signal transduction. These genes involve in intracellular signaling cascade, cell surface receptor linked signal transduction, regulation of signal transduction and other signal transductions (Figure 5.17).

Among the 53 intracellular signaling cascade genes (Figure 5.17a), some of them involve in various protein kinase cascades. Specifically, *Tnf* (tumor necrosis factor) and *Tlr4* (toll-like receptor 4 gene) involve in I-kappaB kinase/NF-kappaB cascade; *Socs1* and *Bsf3* (cardiotrophin-like cytokine factor 1) function in JAK-STAT cascade;, *Spred1* (sprouty protein with EVH-1 domain 1 related sequence), *Sod1*,(superoxide dismutase 1), *Spag9* (sperm associated antigen 9), *Gadd45g* (growth arrest and DNA-damage-inducible 45 gamma) and *Dok2* (docking protein 2) function in MAPKKK cascade. On the other hand, Tbl3 (transducin (beta)-like 3), Gnas (RIKEN cDNA A930027G11), Ptger4 (prostaglandin E receptor 4 (subtype EP4)), Gnat2 (guanine nucleotide binding protein, alpha transducing 2), Gnaq (guanine nucleotide binding protein, alpha q polypeptide) and

Edg3 (endothelial differentiation, sphingolipid G-protein-coupled receptor 3) involve in G-protein coupled receptor protein signaling pathway. Additionally, RAS oncogene family member Ran, Rab11a, Arl1, Arl4, Arl8 (ADP-ribosylation factor-like 1, 4, 8), Arf2, Arf4, Arf6 (ADP-ribosylation factor 2, 4, 6), Sos1 (Son of sevenless homolog 1) and Dnaja3 (DnaJ (Hsp40) homolog, subfamily A, member 3) function in small GTPase mediated signal transduction; Kras (v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog), Rras2 (related RAS viral (r-ras) oncogene homolog 2), Hras1 (Harvey rat sarcoma virus oncogene 1) and Dok2 (docking protein 2) involve in Ras protein signal transduction; Arhgap5 (Rho GTPase activating protein 5), Mcf2l (mcf.2 transforming sequence-like) and Ctnnal1 (catenin (cadherin associated protein), alpha-like 1) function in Rho protein signal transduction; Hspa5 (heat shock 70kD protein 5) and Atp2a2 (ATPase, Ca++ transporting, cardiac muscle, slow twitch 2) are related to ER-nuclear signaling pathway. Other intracellular cascade genes include Inpp4b and Cd8a which are related to second-messenger-mediated signaling, and 17 genes with various functions (Figure 5.17a).

There are 35 genes function in cell surface receptor linked signal transduction (Figure 5.17b). Specifically, Tnfrsf8 (tumor necrosis factor receptor superfamily, member 8), Socs1 (suppressor of cytokine signaling 1) and Rqcd1 (rcd1 (required for cell differentiation) homolog 1) function in cytokine and chemokine mediated signaling pathway; Strap (serine/threonine kinase receptor associated protein), Hbegf (heparin-binding EGF-like growth factor) and Bambi (BMP and activin membrane-bound inhibitor, homolog) function in growth factor receptor signaling pathway; Ptpre, Ptprg, Ptprj (protein tyrosine phosphatase, receptor type E, G, J), Epha2 (Eph receptor A2) and Dok2 (docking protein 2) involve in transmembrane receptor protein signaling pathway through protein amino acid phosphorylation or dephosphorylation; Itga6, Itga9, Itgav (integrin alpha 6, 9, V) and Itgb4bp (integrin beta 4 binding protein) are related to integrin-mediated signaling pathway.

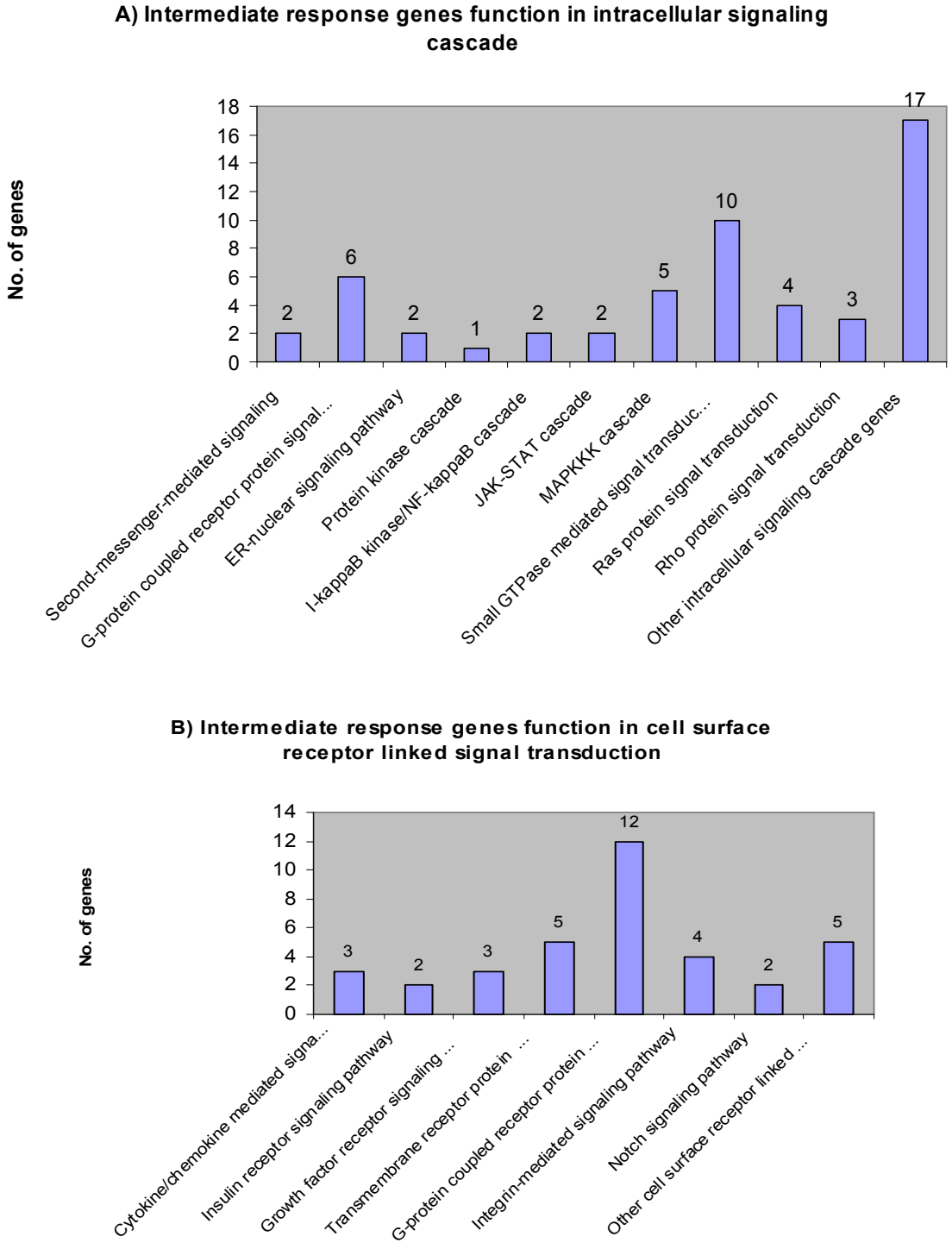


Figure 5.17. Intermediate response genes function in signal transduction. A: above, genes involve in intracellular signaling cascade. B: below, genes function in cell surface receptor linked signal transduction. The number on top of each bar is the number of genes involved in that function.

In addition, Tbl3 (transducin (beta)-like 3), Rgs18 (regulator of G-protein signaling 18), P2ry14 (purinergic receptor P2Y, G-protein coupled 14), Ptger4 (prostaglandin E receptor 4 (subtype EP4)), Gnas, Gnat2, Gnaq, Gpr65, Gpr85, Gpr46 (G-protein coupled receptor 65, 85, 146), Edg3 (endothelial differentiation, sphingolipid G-protein-coupled receptor 3) and Cysltr1 (cysteinyl leukotriene receptor 1) have function in G-protein coupled receptor protein signaling pathway. In contrast, fewer genes involve in other signal transduction pathways, Wdr12 (WD repeat domain 12) and Jag1 (jagged 1) involve in notch signaling pathway, Socs1 and Zfp106 function in insulin receptor signaling pathway, and Il4ra, Cd8a, Bsf3, Hax1 and Fgfrl1 function in other cell surface receptor linked signal transduction. Finally, Tnf, Cish, Socs1, Socs2, Socs6, Cd8a, Jag1 and Spry4 (sprouty homolog 4), Strap (serine/threonine kinase receptor associated protein), Rgs18 (regulator of G-protein signaling 18) as well as Pak1ip1 (PAK1 interacting protein 1) function in regulation of signal transduction, and rest 16 genes have function not included in the those described above, such as cyclin D3, Tnfrsf21 and Tnfrsf11b.

**5.4.3.4. Transcription.** Out of the129 gene (represented by 182 probes) involve in gene transcription, 97 of them encode transcription factors, such as Bcl3 (B-cell leukemia/lymphoma 3), Atf4 (activating transcription factor 4), Arnt2 (aryl hydrocarbon receptor nuclear translocator 2), Atf3, Atf4 (activating transcription factor 3, 4), Rpo1-2, Rpo1-4 (RNA polymerase 1-2, 1-4) and Hnrpr, HnrpD, Hnrpab (heterogeneous nuclear ribonucleoprotein R, D, A/B), 6 zinc finger protein genes (Zfx, Zik1, Zfp91, Zfp451, Zfp238 and Zfp146) and 6 minichromosome maintenance genes (Mcm2, Mcm3, Mcm4, Mcm5, Mcm6, Mcm7). Most of these transcription factors are in nucleus and involve in DNA binding, thus function in regulation of gene transcription.

**5.4.4. Late response genes** (436 probes)

GO biological processes of 243 probes from 436 probes in this category are characterized, representing 203 unique genes. Only a few genes, beta-2 microglobulin (B2m), UL16 binding protein 1 (Ulbp1) and casitas B-lineage lymphoma b (Cblb), involve immune response. The

number of genes as well as the percentage of genes involved in signal transduction and transcription versus all the genes in the same GO category are also relatively low, however, the percentage of genes involved in cell cycle and cell division is fairly high (Figure 5.8).

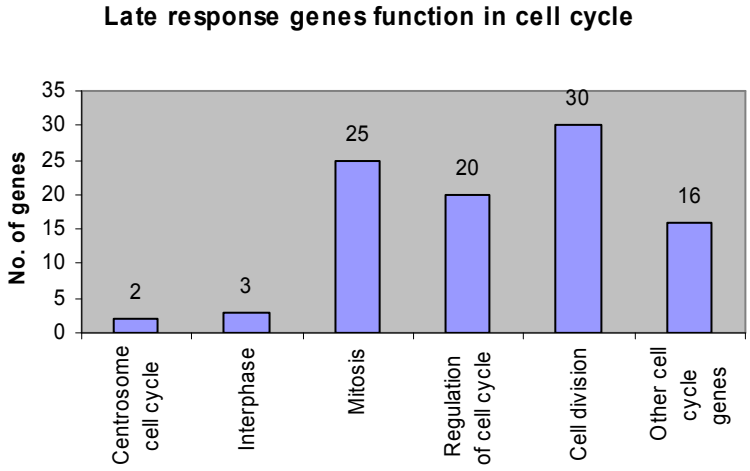


Figure 5.18. Late response genes function in cell cycle. The number on top of each bar is the number of genes involved in that function.

**5.4.4.1. Cell cycle.** Approximately 25% (57 genes) of the unique late response genes function in various phases of the cell cycle (Figure 5.18). In interphase, Mtbp (Mdm2, transformed 3T3 cell double minute p53 binding protein) and expressed sequence AA545217 control the start of cell cycle, while Rbbp8 (retinoblastoma binding protein 8) controls G1/S transition. As expected, most of the cell cycle genes have function related to mitosis, cell division, or regulation of cell cycle, indicating that mouse T cells are ready for proliferation after IL-2 stimulation for 8 hours. Specifically, TPX2 (microtubule-associated protein) and Stmn1 (stathmin 1) function in spindle organization and biogenesis; Brrn1 (barren homolog), Smc4l1 (SMC4 structural maintenance of chromosomes 4-like 1), Nek2 (NIMA (never in mitosis gene a)-related expressed kinase 2), Nusap1 (nucleolar and spindle associated protein 1), Mad2l1 (MAD2 (mitotic arrest deficient, homolog)-like 1) and Cenph (centromere autoantigen H) have function related to chromosome condensation and/or segregation; cyclin B1 gene and Ube2c (ubiquitin-conjugating enzyme E2C) have function related to cytokinesis and regulation of cell cycle. Additionally, cyclin B1, Nek2, Bub1 (budding uninhibited by benzimidazoles 1 homolog), Bub1b, Cdc2a (cell division cycle 2 homolog A) and Cdk2 (cyclin-dependent kinase 2) involve in protein amino acid

phosphorylation, cyclin A2, Kntc1 (kinetochore associated 1), Incenp (inner centromere protein), Fbxo5 (F-box only protein 5), Rcc2 (regulator of chromosome condensation 2), Ccdc5 (coiled-coil domain containing 5), Cdca5 (cell division cycle associated 5), Cdc20 (cell division cycle 20 homolog) and Calmbp1 (calmodulin binding protein 1) function in mitosis or cell division.

As for the 20 genes with function related to regulation of cell cycle (Figure 5.17), cyclin A2, cyclin B1, Cdc2a, Mad2l1, Nusap1 and Ube2c; gene Ris2 (retroviral integration site 2), Clspn (claspin homolog), Cdc45l (cell division cycle 45 homolog-like) and Gmnn (geminin) function in DNA replication, while Brca1 (breast cancer 1) and Msh2 (mutS homolog 2) involve in DNA repairment. Additionally, 16 other cell cycle genes have function related to DNA replication or repair, such as Mcm8 (minichromosome maintenance deficient 8, replication initiation), Lig1 (ligase I, DNA, ATP-dependent), Fanca (Fanconi anemia, complementation group A), Chaf1b (chromatin assembly factor 1, subunit B (p60)), Topbp1 (topoisomerase (DNA) II beta binding protein) and Uhrf1 (ubiquitin-like, containing PHD and RING finger domains, 1). Additionally, Terf2 (telomeric repeat binding factor 2) maintains telomere, Smc2l1 (SMC2 structural maintenance of chromosomes 2-like 1) involves in chromosome condensation and segregation, Brca1 (breast cancer 1) and Kif11 (kinesin family member 11, involves in centrosome separation) function in centrosome cell cycle.

Among the 30 genes with function in cell division, 25 of them also involve in interphase, mitosis and cell cycle regulation. The rest cell division genes either function in cytokinesis (Anln: anillin, actin binding protein, and Prc1: protein regulator of cytokinesis 1), DNA topological change (Top2a: topoisomerase (DNA) II alpha), microtubule-based movement (Kif20a: kinesin family member 20A) or cell division (Cdca3 : cell division cycle associated 3).

**5.4.4.2. Signal transduction.** In contrast to many early (210 genes) and intermediate (96 genes) response genes involve in signal transduction, there are 17 late response genes with function in various signal transduction pathways (Figure 5.19). Specifically, Cblb (Casitas B-lineage lymphoma b) negatively regulates T cell receptor signaling pathway, Rasd2 (RASD family, member 2) and Rhebl1 (Ras homolog enriched in brain like 1) function in small GTPase mediated signal transduction, others like ect2 oncogene, stathmin 1, Racgap1 (Rac GTPase-

activating protein 1), Ddit3 (DNA-damage inducible transcript 3), Depdc1a (DEP domain containing 1a) and Tec (cytoplasmic tyrosine kinase, Dscr28C related) also function in intracellular signaling cascade. Additionally Chaf1b (chromatin assembly factor 1, subunit B (p60)) and Gpr19 (G protein-coupled receptor 19) involve in G-protein coupled receptor protein signaling pathway; Eif4ebp1 (eukaryotic translation initiation factor 4E binding protein 1) functions in insulin receptor signaling pathway; Tle3 (transducin-like enhancer of split 3), Ptch1 (patched homolog 1) and Nfkbia (nuclear factor of kappa light chain gene enhancer in B-cells inhibitor, alpha) involve in other cell surface receptor linked signal transduction pathways, Cblb, Nfkbia and Ptch1 (patched homolog 1) also have function in regulation of signal transduction.

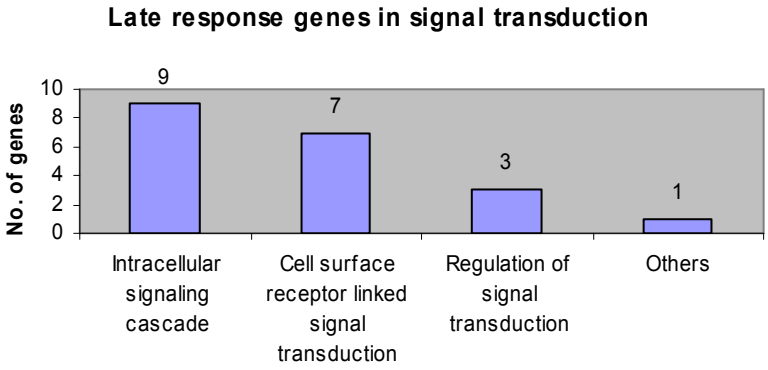


Figure 5.19. Late response genes function in signal transduction. The number on top of each bar is the number of genes involved in that function.

**5.4.4.3. Transcription.** There are also limited number of late response genes (43 genes represented by 49 probes) function in transcription. Totally 30 genes encode transcription factors, majority of them involve DNA binding, such as Zfp52 (zinc finger protein 52) and Dnmt1 (DNA methyltransferase (cytosine-5) 1), only Nfatc2ip (nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 2 interacting protein) and Rbl1 (retinoblastoma-like 1 (p107)) involve in protein binding. Additionally, the function of gene Cbx2 (chromobox homolog 2), Chd4 (chromodomain helicase DNA binding protein 4) and Ing3 (inhibitor of growth family, member 3) is related to chromatin modification.

## 5.5. CONCLUSION

Using a 12 time point series microarray data of murine T cells comprising two initial conditions (IL-2 starved, and IL-2 stimulated), we have inferred networks of activation inhibition relationships between clusters of genes having similar expression profiles. While studying the dynamics of the inferred networks, we found that for the IL-2 stimulated condition, the gene expression profiles of the T cell fluctuates in a cyclic manner. We thus concluded that after IL-2 is being added, T cells are alive and follow cell cycle leading to cell division and proliferation. To reach this conclusion we classified our cluster of genes into four categories, early genes (up-regulated at 1 hour), cyclic genes (up regulated at 1 and 16 hours), intermediate genes (up regulated at 2 hours), and late genes (up regulated at 8 hours). We find that early genes activate other early genes and late genes. Intermediate genes activate late genes and inhibit early genes. Late genes activate other late genes and inhibit early and intermediate genes. A systematic analysis of the gene annotations reveals that early genes are primarily involved in immune response and signal transduction. In particular there is a substantial fraction of early genes involve in immune cell activation and the regulation of immune response. Intermediate genes are involved biosynthesis, translation and metabolism (DNA, RNA, amino acids, nucleotide, sulfur,…). Intermediate genes thus appear to correspond to the S-phase of the cell cycle. Late genes and cyclic genes are mostly involved in the M-phase that is mitosis and cell division.

## 5.6. ACKNOWLEDGEMENTS

This work was funded by Sandia Laboratory Directed Research and Development. Some of the work was also funded by the US Department of Energy's Genomics: GTL program (www.doegenomestolife.org) under project, "Carbon Sequestration in Synechococcus Sp.: From Molecular Machines to Hierarchical Modeling," (www.genomes-to-life.org). Sandia is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. We also would like to thanks Arlene Gonzales and Pamela Lane for collecting cells at various time points.

## 5.7. REFERENCES


Brennan, P., J. W. Babbage, B. M. T. Burgering, B. Groner, K. Reif, and D. A. Cantrell. 1997. Phosphatidylinositol 3-kinase couples the interleukin-2 receptor to the cell cycle regulator E2F. Immunity 7:679.

Brennan, P., J. W. Babbage, G. Thomas, and D. Cantrell. 1999. p70s6K integrates phosphatidylinositol 3-kinase and rapamycin-regulated signals for E2F regulation in T lymphocytes. Mol. Cell. Biol. 19:4729.

Evans, G. E., M. A. Goldsmith, J. A. Johnston, W. Xu, S. R. Weiler, R. Erwin, O. M. Z. Howard, R. T. Abraham, J. J. O'Shea, W. C. Greene, and W. C. Farrar. 1995. Analysis of interleukin-2 dependent signal transduction through the Shc/Grb2 adaptor pathway. J. Biol. Chem. 270:28858.

Frearson, J. A., and D. R. Alexander. 1997. The role of phosphotyrosine phosphatases in haematopoietic cell signal transduction. BioEssays 19:417.

Friedmann, M. C., T. S. Migone, S. M. Russell, and W. J. Leonard. 1996. Different interleukin 2 receptor β-chain tyrosines couple to at least two signaling pathways and synergistically mediate interleukin 2-induced proliferation. Proc. Natl. Acad. Sci. USA 93:2077.

Gadina, M., C. Sudarshan, and J. J. O'Shea. 1999. IL-2, but not IL-4 and other cytokines, induces phosphorylation of a 98 kDa protein associated with SHP-2, phosphatidylinositol 3-kinase, and Grb2. J. Immunol. 162:2081.

Gesbert, F., M. Delespine-Carmagnat, and J. Bertoglio. 1998. Recent advances in he understanding of interleukin-2 signal transduction. J. Clin. Immunol. 18:307.

Gilmour, K. C., R. Pine, and N. C. Reich. 1995. Interleukin 2 activates Stat5 transcriptional factor (mammary gland factor) and specific gene expression in T lymphocytes. Proc. Natl. Acad. Sci. USA 92:10772.

Gorelik, L., and R. A. Flavell. 2002. Transforming growth factor-β in T-cell biology. Nat. Rev. Immunol. 2:46.

Gu. H., J. C. Pratt, S. J. Burakoff, and B. G. Neel. 1998. Cloning of p97/Gab2, the major SHP-2 binding protein in hematopoietic cells, reveals a novel pathway for cytokine-induced gene activation. Mol. Cell 2:729.

Gu, H., H. Maeda, J. J. Moon, J. D. Lord, M. Yoakim, B. H. Nelson, and B. G. Neel. 2000. New role for Shc in activation of the phosphatidylinositol 3-kinase/Akt pathway. Mol. Cell. Biol. 20:7109.

Harmer, S. L., and A. L. DeFranco. 1997. Shc contains two Grb2 binding sites needed for efficient formation of complexes with SOS in B lymphocytes. Mol. Cell. Biol. 17:4087.

Higuchi, M., H. Asao, N. Tanaka, K. Oala, T. Takeshita, M. Nakamura, J. Van Snick, and K. Sugamura. 1996. Dispensability of Jak1 tyrosine kinase for interleukin-2-induced cell growth signaling in a human T cell line. Eur. J. Immunol. 26:1322.

Hill, C. S., and R. Treisman. 1995. Transcriptional regulation by extracellular signals: mechanisms and specificity. Cell 80:199.

Hou, J., U. Schinder, W. J. Henzel, S. C. Wang, and S. C. McKnight. 1995. Identification and purification of human stat proteins activated in response to interleukin-2. Immunity 2:321.

Jain, A. K., Murty, M.N., and Flynn, P.J. 1999. Data Clustering: A Review. ACM Computing Surveys, Vol 31, No. 3, 264-323.

John, S., C. M. Robbins, and W. J. Leonard. 1996. An IL-2 response element in the human IL-2 receptor a chain promoter is a composite element that binds Stat5, Elf-1, HMG-1(Y), and GATA family protein. EMBO J. 15:5627.

Johnston, J. A., M. Kawamura, R. A. Kirken, Y. Chen, T. B. Blake, K. Shibuya, J. R. Ortaldo, D. W. McVicar, and J. J. O'Shea. 1994. Phosphorylation and activation of the Jak-3 Janus kinase in response to interleukin-2. Nature 370:151.

Kauffman, S. A. 1969. Metabolic Stability and Epigenesis in Randonly Constructed Genetic Nets. J. Theor. Biol. 22: 437.

Lecine, P., M. Algarte, P. Rameil, C. Beadling, P. Bucher, M. Nabholz, and J. Imbert. 1996. Elf-1 and Stat5 bind to a critical element in a new enhancer of the human interleukin-2 receptor a gene. Mol. Cell. Biol. 16:6829.

Letterio, J. J., and A. B. Roberts. 1998. Regulation of immune responses by TGF-β. Annu. Rev. Immunol. 16:137.

Basso, K, Margolin, A. A., Stolovitzky, G., Klein, U. Dalla-Favera, R., and Califano, A. 2005. Reverse engineering of regulatory networks in human B cells, Nature Genetics 37: 382.

Lin, J. X., and W. J. Leonard. 2000. The role of Stat5a and Stat5b in signaling by IL-2 family cytokines. Oncogene 19:2566.

Lord, J. D., B. C. McIntosh, P. D. Greenberg, and B. H. Nelson. 1998. The IL-2 receptor promotes proliferation, bcl-2 and bcl-x induction, but not cell viability through the adapter molecule Shc. J. Immunol. 161:4627.

Lord, J. D., B. C. McIntosh, P. D. Greenberg, and B. H. Nelson. 2000. The IL-2 receptor promotes lymphocyte proliferation and induction of the c-myc, bcl-2, and bcl-x genes through the trans-activation domain of Stat5. J. Immunol. 164:2533.

Matikainen, S., T. Sareneva, T. Ronni, A. Lehtonen, P. J. Koskinen, and I. Julkunen. 1999. Interferon-a activates multiple STAT proteins and upregulates proliferation-associated IL-2Ra, c-myc, and pim-1 genes in human T cells. Blood 93:1980.

Matsumara, I., T. Kitamura, H. Wakao, H. Tanaka, K. Hashimoto, C. Albanese, J. Downward, R. G. Pestell, and Y. Kanakura. 1999. Transcriptional regulation of the cyclin D1 promoter by STAT5: its involvement in cytokine dependent growth of hematopoietic cells. EMBO J. 18:1367.

Matsumura, I., J. Ishikawa, K. Nakajima, K. Oritani, Y. Tomiyama, J. Miyagawa, T. Kato, H. Miyazaki, Y. Matsuzawa, and Y. Kanakura. 1997. Thrombopoietin induced differentiation of a human megakaryoblastic leukemia cell line, CMK, involves transcriptional activity of p21WAF1/Cip1 by Stat5. Mol. Cell. Biol. 17:2933.

Moon, J. J., and B. H. Nelson. 2001. Phosphatidylinositol 3-kinase potentiates, but does not trigger, T cell proliferation mediated by the IL-2 receptor. J. Immunol. 167:2714.

Moriggl, R., D. J. Topham, S. Teglund, V. Sexl, C. McKay, D. Wang, A. Hoffmeyer, J. van Deursen, M. Y. Sangster, K. D. Bunting, et al. 1999. Stat5 is required for IL-2 induced cell cycle progression of peripheral T cells. Immunity 10:249.

Mui, A. L., H. Wakao, T. Kinoshita, T. Kitamura, and A. Miyajima. 1996. Suppression of interleukin-3-induced gene expression by a C-terminal truncated Stat5: role of Stat5 in proliferation. EMBO J. 15:2425.

Nelson, B. H., and D. M. Willerford. 1998. Biology of the interleukin-2 receptor. Adv. Immunol. 70:1 (In Advances in Immunology, Vol. 70. Academic Press, New York, p. 1).

Ravichandran, K. S., U. Lorenz, S. E. Shoelson, and S. J. Burakoff. 1995. Interaction of Shc with Grb2 regulates association of Grb2 with mSOS. Mol. Cell. Biol. 15:593.

Ravichandran, K. S., V. Igras, S. E. Shoelson, S. W. Fesik, and S. J. Burakoff. 1996. Evidence for a role for the phosphotyrosine binding domain of Shc in interleukin-2 signaling. Proc. Natl. Acad. Sci. USA 93:5275.

Schaeffer, H. J., and M. J. Weber. 1999. Mitogen-activated protein kinases: specific messages from ubiquitous messengers. Mol. Cell. Biol. 19:2435.

Schindler, C. 1999. Cytokines and JAK-STAT signaling. Exp. Cell Res. 253:7.

Smola, A.J. and Scholkopf, B. 1998 A tutorial on support vector regression. NeuroCOLT2 Technical Report NC2-TR-1998-030.

Takeshita, T., K. Ohtami, H. Asao, S. Kumaki, M. Nakamura, and K. Sugamura. 1992. An associated molecule, p64 with IL-2 receptor b chain: its possible involvement in the formation of the functional intermediate IL-2 receptor complex. J. Immunol. 148:2154.

Trefethen, Llyod and Bau, David. 1997. Numerical Linear Algebra. SIAM.

Witthuhn, B. A., O. Silvennoinen. O. Miura, K. S. Lai, C. Cwik, E. T. Liu, and J. N. Ihle. 1994. Involvement of the Jak-3 Janus kinase in signaling by interleukins 2 and 4 in lymphoid and myeloid cells. Nature 370:153.

DISTRIBUTION:

| | | |
|---|---|---|
| 1 | 0310 | Danny Rintoul, 1410 |
| 1 | 0310 | Elebeoba May, 1412 |
| 1 | 0310 | George Davidson, 1400 |
| 1 | 0310 | Shawn S. Martin, 1412 |
| 1 | 0829 | Ed Thomas, 12337 |
| 1 | 0895 | Anthony Martino, 8332 |
| 1 | 0895 | David Haaland, 8332 |
| 1 | 0895 | Jeri Timlin, 8332 |
| 1 | 1110 | Bob Carr, 1415 |
| 1 | 1110 | Alex Slepoy, 1435 |
| 1 | 1411 | Mike Sinclair, 1824 |
| 1 | 1413 | Grant S. Heffelfinger, 8330 |
| 3 | 1413 | Jean-Loup Faulon, 8333 |
| 1 | 9291 | Zhaoduo Zhang, 8321 |
| 1 | 9292 | Malin Young, 8321 |
| 1 | 9018 | Central Technical Files, 8945-1 |
| 2 | 0899 | Technical Library, 9616 |
| 1 | 0123 | D. Chavez, LDRD Office, 1011 |